A conceptionary for speech and hearing in the context of machines and experimentation

Technical Report · January 1978

DOI: 10.13140/RG.2.2.24303.28329

David Hill
The University of Calgary

THE UNIVERSITY OF CALGARY

A CONCEPTIONARY FOR SPEECH AND HEARING IN THE CONTEXT OF MACHINES AND EXPERIMENTATION

by

D.R. Hill

Research Report Number 78/27/6

January 1978
(Revised 2001, 2018)

DEPARTMENT OF
COMPUTER SCIENCE

A CONCEPTIONARY FOR SPEECH AND HEARING IN THE CONTEXT OF MACHINES AND EXPERIMENTATION

by

D.R. Hill

Research Report Number 78/27/6

January 1978
(Revised 2001, 2018)

A Conceptionary for Speech & Hearing in the Context of Machines and Experimentation

by

David R. Hill
hilld@ucalgary.ca

Introduction

A conceptionary, like a dictionary, is useful in learning the meanings of words. Unlike a dictionary, it is designed to make the reader work a little, and to develop associations and a conceptual framework for the subject of the conceptionary. It is not an encyclopaedia, though it has as one function, in new or inter-disciplinary areas, the drawing together and reconciliation of disparate sources. Like Marmite (a yeast extract), it is designed to be nutritious and highly concentrated. Do not be put off by the consequent strong flavour. The aim is to achieve broad coverage of material from areas such as psychology and physics, as well as from speech recognition and synthesis, because such material is highly relevant to speech researchers and often difficult to track down. Historical material is included for the same reasons. The view of the field expressed is a rather personal view. The author wishes to thank Mr. R.L. Jenkins for his careful reading of the manuscript and many useful comments and corrections, while laying sole claim to any errors, omissions, or lack of clarity that remain.

Note that although URLs that allow access to some of the major speech research centres are included within this document, many more centres exist. Those that are provided are those that were active at the time the document was originally prepared. A search using the *Google* search facility, with the phrase "speech communication research labs" will turn up a large selection of these, but not necessarily including any of those provided below!

Please email any comments, corrections or suggestions to hilld@.ucalgary.ca. While every effort has been made to present accurate information, it is not intended to support safety critical systems. Material should always be cross-checked with other sources.

Concepts, definitions, etc.

AH1/AH2

Amplitudes of two kinds of hiss used as part of the excitation arrangements of the Parametric Artificial Talker (PAT—the original resonance analog speech synthesiser). Hiss 1 is fed into the formant branch of the filter function, and represents aspiration noise, Hiss 2 is fed to a separate filter and represents the noises (such as the /s/ in sun and the /sh/ in shun) produced by forcing air through a narrow constriction towards the front of the oral cavity, and therefore assumed to be relatively unaffected by the usual oral resonances, or formants.

Ax /Ao

Amplitude of voiced energy excitation (often called larynx amplitude) injected into the filter function of a speech synthesiser. Corresponds to voicing effort in natural speech.

absolute refractory period
See *nerve cell*

acoustic analog
A construct that models the frequency-time-energy pattern of some sound without regard to how it is produced, or the mechanism involved. Such a model is only constrained by the bandwidths and dynamic range inherent in its structure. See *also resonance analog, physiological analog* and *terminal analog*.

acoustic correlates
Those acoustic phenomena that correlate (or necessarily co-occur) with articulatory events. If they can be discovered and well defined, they may be used to make inferences about articulation. For example, a burst of noise usually occurs when a blockage in the vocal passages is released, so the detection of a burst of noise is some evidence that the input speech to a recogniser contained a stop sound. Unfortunately, the acoustic correlates of articulatory events are not well-defined. Speech recognisers require means of normalisation, prediction, extrapolation, and the like.

acoustic domain
Description or existence in terms of the amount of energy present at given times and frequencies. A spectrogram is an acoustic domain description of speech. It is contrasted with time domain descriptions. A time domain description of speech is the time waveform.

acoustic intensity
Is defined as the average power transmitted per unit area in the direction of wave propagation. I = p(rms)2/r.c watts/square metre where p(rms) is the root-mean-square effective acoustic pressure, r is the density of the medium (air at 20°C and standard pressure gives r = 1.21 kg/cubic metre), and c the speed of sound in the medium (same air gives c=343 meters/sec).

acoustic intensity level
See *sound intensity level*

acuity
The power of discrimination in some sensory mode. See *difference limen*.

afferent
Incoming from the periphery to the central nervous system. Thus afferent nerves are sensory nerves running in from the sensory receptors scattered around an organism.

affricate
A speech sound formed by the close juxtaposition of a stop and a fricative, in that order. For English, voiceless and voiced examples occur at the start of “chips” and “judge” respectively. The affricate is distinguished from the corresponding stop-fricative combination chiefly by the much shorter burst of friction noise which accompanies the explosive separation of the articulators (thus my chip is different from might ship, a problem in notation solved by introducing a hyphen to distinguish the juncture).

AGC
See *automatic gain control*

allophone
see *phoneme*

Ames room
A room of unusual shape, designed by Professor Ames of Princeton University, so that the retinal image of the room from a particular view point coincides exactly with the image that a normal room would produce. People walking about the room appear to shrink and grow as they occupy larger and smaller parts which are further and nearer (respectively), because the brain assumes the room is normally shaped, and therefore misinterprets the changing image size as size change instead of distance change. Intellectual knowledge does not dispel the illusion. Experience of trying to do things inside the room can cause the room to be perceived as it really is. The effect lends

itself to dramatic demonstration on film, as it depends on viewing with only one eye. See also *the moon illusion*.

alveolar ridge

As the hard palate runs towards the top front teeth from behind, a buttress enclosing the roots of the teeth is encountered. This is called the alveolar ridge.

anacrusis

A term borrowed from music. In music, an anacrusis is a series of notes that are not counted towards the total note duration in a bar—they are "extra". Jassem uses the term anacrusis for what Abercrombie and others have called proclitic syllables—syllables at the end of a rhythmic unit in speech that belong, grammatically speaking, with the following rhythmic unit. Jassem chose the term deliberately because, in his theory of British English speech rhythm, the anacruses are not to be counted in the duration of the rhythmic unit (called a foot by Abercrombie and others or a rhythm unit by Jassem). See *isochrony, foot, proclitic,* and *enclitic*.

articulation

The process of placing the tongue, lips, teeth, velum, pharynx and vocal folds in the state or succession of states required for some utterance. Also used for one such state (e.g. "consonant articulation"). There is unconcious ambiguity in just what is meant in normal use (e.g. "He articulated clearly" implying that articulation itself is "sound", rather than a placing of the articulators in some "articulatory posture". One could easily articulate an isolated consonant, without actually producing sound. Sound only results when excitation energy of some sort is supplied (e.g. by using muscles to expel air from the lungs so that the vocal folds vibrate.

articulation index

A method of estimating speech intelligibility based upon the division of the speech spectrum into 20 bands contributing equally to intelligibility. It is assumed that, for each band, the intelligibility contribution is proportional to the signal to noise ratio in the band, being 100% for s/n greater than or equal to 30 db and zero for s/n less than or equal to 0 db. The original method (French & Steinberg) using 20 bands of equal weight, and of width related to the critical band distribution, has been adapted to fifteen 1/3 octave bands with appropriately differing weights, which more suited to the measuring apparatus that is easily available. Under some circumstances (e.g. when given directional cues or special instructions, etc.) speech may be intelligible even when masked at an overall signal to noise ratio of -10 or even -15 db. This should not be taken as contradiction of the validity of articulation index calculations.

articulatory synthesis

Speech synthesis based on the use of articulatory constraints. Articulatory constraints may be applied to the generation of parameters for any synthesiser, but true articulatory synthesis depends on the articulatory constraints being inherent in the implementation of the synthesis system, as the reproduction of all the details is otherwise very difficult if not impossible. A transmission-line analog with a nasal branch emulates the propagation of sound waves in the acoustic tube formed by the vocal tract directly. Thus the full spectrum is produced, with correct energy interaction for nasal sounds, and correct shapes for the formants. True articulatory synthesis is to other approaches to synthesis as true ray tracing is to polygonal modeling in computer graphics. By modeling the physics of reality directly, articulatory synthesis achieves a high degree of fidelity to nature. The problems arise from a need for a great deal of computational power, the need to manage the control problem successfully, and the need for accurate articulatory data. Ideally, the various energy sources should be emulated on the basis of physics, but the only complete articulatory synthesis system currently available (originally a commercial development by Trillium Sound Research for the NeXTSTEP) operating system, is now available under a General Public Licence from the the Free Software Foundation https://www.gnu.org/software/gnuspeech/). The system does not model the fricative noise from constrictions, but injects appropriate waveforms at appropriate places in the vocal tract, because of a lack of computational power, as well as lack of suitable models. The system includes English text-to-speech generation The research is continuing. See the paper *Low-level articulatory synthesis: A working text-tospeech solution and a linguistic tool* (https://doi.org/10.1017/cnj.2017.15) for more detail.

artificial ear

Not a prosthetic device (yet). A device that simulates the acoustics of the head and the ear cavity of a human being, measuring (by means of a calibrated transducer) the pressure that would occur at the ear-drum of a real ear. It is used in determining the total effect of earphones plus cushions on the fidelity of reproduction of speech without resorting to sophisticated, highly trained (i.e. expensive) real listeners for subjective testing. The latter is a far more reliable method of evaluation. See *pinna*.

artificial larynx

A device used for prosthesis in laryngectomized persons. It is held against the throat and, when activated by a button, it injects vibratory energy into the vocal tract as a substitute for the lost voice capability. It is not too difficult to use, but less convenient than oesophagal speech.

artificial mouth

Not a prosthetic device (yet). A model designed to simulate the acoustic characteristics of the head and mouth upon the radiated sound. Used, for example, in evaluating microphones.

ARU

See *audio response unit*.

aspiration

The escape of breath through a relatively unconstricted vocal tract, without accompanying vibration of the vocal folds. Some turbulent or friction noise is generated but it is not clear whether this is due to turbulence at the glottis or along the walls of the vocal tract, in general.

ASR

See *automatic speech recognition*

audibility

The degree to which an acoustic signal can be detected by a listener. See also *intelligibility*.

audio response unit

A peripheral device attached to a computer to allow voice messages to be sent to users connected by telephone. A Touch-Tone phone is used in most current applications to provide input. The ARU may replay direct recordings, or compressed recordings, or it may synthesise the speech using rules based on general speech knowledge. The commonest current method uses compressed recordings based on Linear Predictive Coding (LPC) parameters.

audiogram

See *audiometry*

audiometry

The process of determining the frequency distribution of a subject's absolute threshold of hearing (in db, compared to "normal" threshold). Pure tone stimuli at selected frequencies are used to sample this distribution by presenting them at various intensities and noting response/no-response from the patient. Quiet conditions (noise not greater than 10 db below threshold of hearing in the booth), noise shielded earphones, and special instruments are used. Each ear is evaluated separately. The plot of threshold against frequency for the two ears is called an audiogram, and shows threshold in db (referenced the normal threshold) as a function of frequency.

audiometer

An instrument calibrated to produce tones and other sounds at known intensities. It is designed to produce audiograms for subjects whose hearing is to be tested.

auditory cortex

That part of the cortex (the convoluted outer layer of the brain, representing the highest degree of evolution) that is primarily responsible for processing auditory signals (i.e. those nerve impulses originating from the ear—specifically from the basilar membrane).

auditory meatus

In full, the external auditory meatus. The canal connecting the eardrum to the outside air. The external opening of the meatus is within the area of the pinna. (The pinnae are the things we are talking about when we say "My ears are cold"). The meatus is slightly curved (which prevents the ear-drum being viewed without instruments - e.g. an otoscope); it is about 3 cm long and about 1 cm or less in diameter. Hairs and wax glands protect against ingress of insects etc. Excess wax can cause some loss of hearing. See also *pinna*.

auditory nerve
See *modiolus*

auditory pathways
The paths traced by the nerves leading from the organ of Corti in the cochlea to the auditory cortex. In logical order, the significant staging posts are: organ of Corti: Corti ganglion: ventral and dorsal cochlear nucleii: superior olive: inferior colliculus: medial geniculate: and finally auditory cortex. The signals are fed roughly equally to both halves of the brain (cross-overs occuring at the level of medulla and the path between there and the inferior colliculus): also side branches occur in the neighbourhood of the medulla (the medulla embraces the trapezoid body, the dorsal and ventral cochlear nucleii, and the two superior olives), at the lateral lemniscate and nucleus, and at the cerebellar vermis. The dorsal cochlear nucleus sends all its outgoing processes to the contralateral side of the brain, though there are connections back in the inferior colliculus region. Based on their neurophysiological studies of cats, Whitfield and Evans (Keele University, U.K.) have stated that the primary function of the auditory cortex is the analysis of time pattern rather than stimulus frequency.

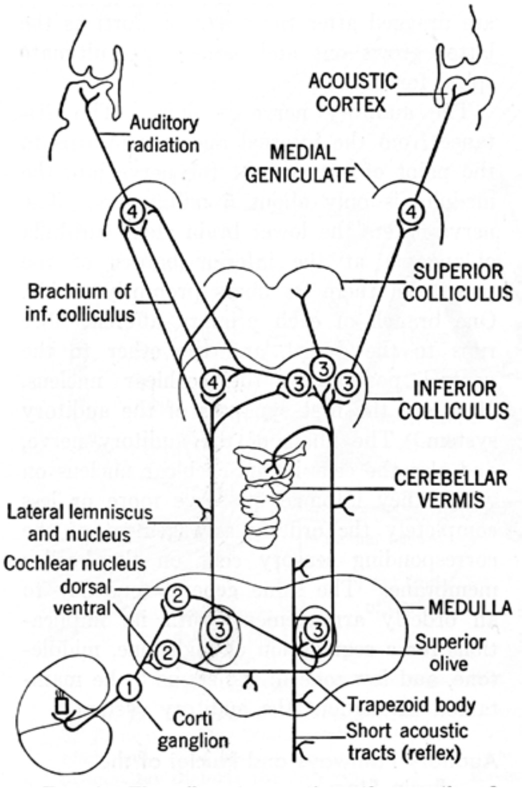

Figure 1: *Afferent acoustic pathways*

From Handbook of Experimental Psychology, p1120
S.S. Stevens (ed.). © 1951 John Wiley & Sons.
Used with permission of John Wiley & Sons.

automatic gain control
A device often included in electronic audio processing systems to limit the amplitude of the signal without actually clipping it. A long-term average of the speech energy is taken and used to control the gain of the amplifier to the system, turning the gain down as the input speech energy increases. Thus signal levels may be kept near optimum levels without overload. If required, the control signal may be transmitted along with the compressed signal to allow re-expansion to the original dynamic range. Such a device is called a compander (naturally). Compression involves distortion as well. The time constants involved in deriving the control signal are important. Faster action is more likely to prevent overload, but produces greater distortion. Longer time constants may be used where the average signal is not expected to vary quickly. See also *clipped speech*.

automatic speech recognition
The process of automatically (that is, by machine) rendering spoken sounds as correct language symbols. It has been contrasted with automatic speech understanding which means the production of a correct response by a machine for given verbal input. The problem of recognising unlimited vocabulary for arbitrary speakers in continu-

ous speech remains an unsolved goal. Most current systems require some degree of training, recognise relatively restricted vocabularies, and only deal with isolated words or phrases (IWR). Continuous Speech Recognition (CSR) is necessary for normal speech as there are no breaks or boundaries between words in normal utterances, unless a pause occurs. It is quite difficult to speak in such a way that a pause is inserted before every word.

Ax

The amplitude of glottal excitation supplied to a synthesis model (the "larynx amplitude").

axon

The output process of a nerve cell (a process is an outgrown limb or filament).

bandwidth

That contiguous band of frequencies that may be transmitted through some signal processing system with no more than 3 db loss of power (that is, the energy in no frequency within the band is attenuated to less than half the original power). A signal is often loosely ascribed a "bandwidth". This really refers to the bandwidth of the processing system that would be needed to process the signal without introducing unacceptable distortion or loss of significant components. For complete fidelity, the bandwidth of the processing system should at least equal the frequency spectrum range of the signal's significant components.

bar

In the sense of "unit of measurement" a bar is the pressure exerted by the "standard atmosphere". Unfortunately, as with early temperature scales, a slight miscalculation has meant that the standard atmosphere accepted today actually exerts a pressure of 1029 millibars (1.029 bars). In modern terms: 1 bar = 0.1 MPa (MegaPascals) where a Pascal is 1 nt/m2 (Newton per Square Metre), and is the standard SI unit of pressure.

basilar membrane

The cochlea, or organ of hearing, is a spiral chamber. It is divided, longitudinally, into two thinner tubes (the scala vestibuli and scala tympani) by the cochlear partition. The cochlear partition is itself a duct, triangular in cross section, bounded on the shortest side by the outer wall of the cochlear wall, by Reissner's membrane, and—on the third side—a combination of a bony shelf and the basilar membrane. The bony shelf is cantilevered out from the modiolus, or centre-post, of the spiral chamber formed by the cochlea. The bony shelf starts wide near the oval and round windows at one end of the cochlear spiral, and narrows down towards the helicotrema, an opening at the other end of the cochlea's spiral tube that joins the two perilymph-filled scala. Thus the basilar membrane starts off narrow by the windows, and becomes progressively wider towards the helicotrema, since the tube forming the cochlea is fairly constant in diameter. The basilar membrane is about 3.5 cm long and varies in width from 0.1 mm at the window end to 0.5 mm at the helicotrema end, making about 2.5 turns in the spiral. The scala are about 2 mm in diameter throughout. Bekesy remarks sanguinely that "it is difficult to carry out experiments with dimensions as small as these." The basilar membrane bears, on its inner surface, the organ of Corti, which activates the sensitive hair cells and hence generates nerve pulses to the brain. See also, *cochlea*, *tectorial membrane* and *membranous labyrinth*.

BBN

Bolt, Beranek, and Newman. An important company in speech, linguistics, and computers. Beranek, one of the founders, is prominent in the early acoustics research literature.

Bell Telephone Laboratories

Bell Telephone Laboratories, better known as Bell Labs, was perhaps the best known research establishment in the world with leading edge research in a mind-boggling number of areas, originally funded by profits from telephone network operation. Formed in 1907 by the combination of AT&T and Western Electric Engineering Departments, but not named till 1925. A consent decree split AT&T in 1956, but Bell Labs continued as one of the world's top two or three speech and communication research laboratories. In 1996, AT&T underwent another major re-organisation and splitting to meet changing world market conditions. Bell Laboratories now forms part of Lucent Technologies. See the relevant Lucent Technologies history page on their web site. A web site specific to Bell Labs is also available. Unix was invented there by Dennis Ritchie and Ken Thompson; James Flanagan, of speech research fame, was there; Claude Shannon, of information theory fame, was there. The sound spectrograph was invented there by H.K. Dunn. The first work by Homer Dudley and his colleagues on vocoders and speech communication was carried out there. The list goes on. The Bell Systems Technical Journal provides a major collection of papers on their research. In 1996 AT&T the parent company, spun off Bell Laboratories, along with most of its equipment manufacturing business, into a new company named Lucent Technologies. There were a number of further changes, including a merger with Alcatel to create Alcatel-Lucent labs. In 2008, Alcatel-Lucent announced it was pulling out of basic science, material physics, and semiconductor research, and it would instead focus on more immediately marketable areas, including networking, high-speed electronics, wireless networks, nanotechnology and software. The organisation has since been bought by Nokia,

beats

When two pure tones fairly close together in frequency are sounded together, a listener hears the intensity waxing and waning at a rate that reflects the difference between the two tones precisely. The effect is due to the successive cancellation and reinforcement of one tone by the other as they get in and out of step (in and out of phase). Over part of the range where difference effects occur, many listeners hear a third tone of the appropriate pitch (i.e. the difference frequency).

bilabial

Of, or with, two lips (from the Latin).

binary digit

A representation of two possible states ("1" and "0") which may form part of the representation of a number in base 2 arithmetic. See also *bit*.

binaural

Pertaining to two ears. See *diotic* and *dichotic*.

bit

Strictly speaking, a unit of information. In common parlance, a bit is a binary digit, which is a physical signal (typically in a computer memory, register, etc.). It has been suggested that since the term bit (binary digit) is used in information theory as a fundamental unit of information measure, the term binit (binary digit) should be used for the physical signal. The reason: although a binit can theoretically transmit one bit of information, it can only do so under noiseless (and hence impossibly perfect) conditions. See *information* and *redundancy*.

black box

A device whose detailed internal construction is unknown, but which has some (usually desired) input-output relationship. It originated during the second world war, due to the practice of painting aircraft electronic boxes black and ensuring that the contents were kept secret.

bone conduction

The path of sound transmitted into the cochlea by vibration of the surrounding bone, however vibration of the bone is produced.

bony labyrinth

The passages inside the temporal bone that contain the mechanical parts associated with hearing and balance. See *membranous labyrinth* and *cochlea*

breathy voice

During phonation, the vocal folds vibrate from relatively open to relatively or completely closed. If the folds are not completely closed, air will pass between the folds and inject random energy into the vocal tract, which adds noise or "breathiness" to the overall voicing energy. The small gap that remains during closure is called a "glottal chink", and is characteristic of many female vocalisations ("breathy voice").

BSTJ

The Bell System Technical Journal. An important research journal published by the Bell Labs. Volume 57, Number 6, Part 2 for July-August 1978 was the original technical publication on Unix, describing in a collection of papers work which started in 1969 when Ken Thompson started working on a cast-off PDP-7 computer and Unix was conceived. Shannon's original work on information theory was first published there. An important source of original work on speech communication and processing. See *Bell Telephone Laboratories*.

buzz/hiss switching

That part of a speech bandwidth compression system concerned with detecting and implementing the change in excitation corresponding to the contrast between voicing and fricative noise. In its simplest form a circuit is provided that detects voicing (not an easy or entirely solved problem) and switches the excitation of the synthesiser part to "buzz" (voiced excitation) or "hiss" (random excitation) accordingly. Of course, such a simple system does not produce correct voiced fricatives. See voicing detection.

capacity

The rate at which information (bits) may be transmitted over a defined channel. See also *bit*.

CCRMA

Centre for Computer Research on Music and Acoustics, at Stanford University, original home of Perry Cook, Julius Smith and others. Perry Cook is now at Princeton University.

cells of Claudius

Cells capable of generating D-C potentials and lying on the basilar membrane beyond the outer edge of the organ of Corti.

centre clipping

See *clipped speech*.

cepstral analysis

If the logarithm of the spectral amplitudes produced by a Fourier analysis of a time series is, itself, treated as a time series (sic) and subjected to a second Fourier analysis, the envelope is broken up into the underlying fine structure (pitch-/excitation-determined in the case of voiced speech) and broad envelope characteristics (formant-/vocal-tract-filter-determined in the case of speech). The terms of the new Fourier series are now called rahmonics (instead of harmonics), and the distribution of rahmonics is called a cepstrum which displays the signal in the quefrency domain. Such analysis is very attractive for speech since, in theory, and to a large extent in practice, one can take the cepstrum, remove some component and then carry out an inverse transform back into the frequency domain. Thus, for example, one can remove the effect of the fine structure on the spectrum and leave only the formant envelope. At the same time, the presence of voicing, and its frequency, may be determined from the component removed.

cepstrum

See *cepstral analysis*.

cerebellar vermis

See *auditory pathways*.

channel vocoder

See *vocoder*.

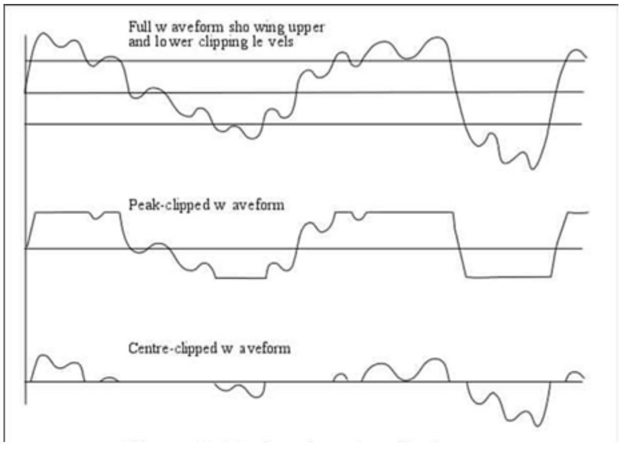

Figure 2: *Peak- and centre-clipped speech*

clipped speech

A speech waveform makes amplitude excursions about some zero level as time progresses. Removal of some of these in various amplitude bands is called "clipping." Peak clipping refers to removal of the largest positive and negative amplitude bands (leaving flattened peaks and troughs not exceeding the largest bands remaining). Centre clipping involves the removal of the centre bands of amplitude. The amount of clipping is expressed in terms of the number of decibels by which the maximum amplitude is reduced: clipping = 20 log (A0/A) db, where the reference is A (the original amplitude) and Ao (the clipped amplitude), but the number of decibels is kept positive by inverting the ratio to keep it greater than 1. Clipping need not always be symmetric about the zero axis, though it usually is. It introduces distortion which may or may not be audible.

coarticulation

The effect of of speech sounds that are fairly close together upon each other. The vowel at the beginning of "abbey" is not the same as the vowel at the beginning of "abu", because coarticulation works backwards as well as forwards, and affects more than just the immediate neighbours. It is due at least in part to the anticipation of articulations to come, and the effects of articulations that are past upon the current posture, and hence upon the sounds that are produced at any given moment when speaking.

cochlea

That part of the bony labyrinth concerned with transducing mechanical pressure waves into nerve signals. The active part is the basilar membrane and associated structures. Its name comes from the Latin meaning "snail" which is what the cochlea spiral resembles. The whole structure is deeply embedded in the temporal bone and exceedingly difficult to work on, either surgically or experimentally. Georg von Bekesy obtained a Nobel prize for his work on hearing, which involved exceedingly delicate dissection, and study of the cochlea and other related structures, by careful ingenious experiments. The space inside the cochlear spiral is a fairly uniform tube of total diameter about 5 mm, 35 mm long, wound around the modiolus (or centre-post) in a decreasing spiral, from the start near the round and oval windows to the end where is found the helicotrema, making roughly 2 1/2 turns. This space is divided longitudinally into two (the scala vestibuli and the scala tympani) by the cochlear partition—itself a fluid filled duct containing among other things the organ of Corti. The cochlear partition is filled with endolymph, while the scala are filled with perilymph, and are joined by the helicotrema. Sound pressure waves are transmitted into the scala vestibuli by the stapes acting on the oval window. The complex hydrodynamic and elastic behaviour of the fluids and structures as the pressure waves affect the fluids on both sides of the cochlear partition, lead to frequency discriminating nerve signals being generated in monotonic progression from low to high frequency along the basilar membrane by the relative movement of the tectorial membrane and the hair cells. The nerve pulses at any given frequency below about 3000 hz tend to occur in phase with the input sound pressure waveform, which is the basis of the "VolleyTheory" of pitch perception. See *basilar membrane.*

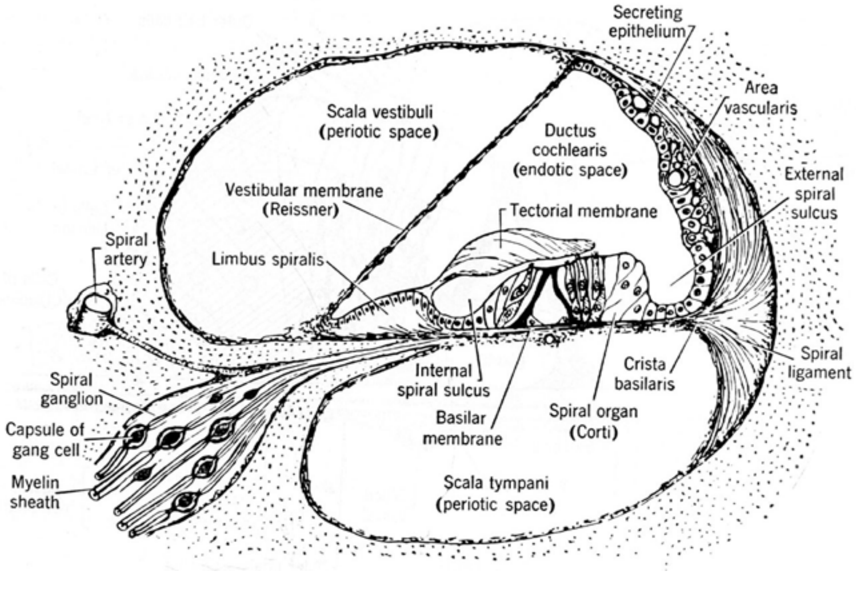

Figure 3: *Cross-section of the cochlear canal*

Cross-section of cochlea canal, 1117, (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

See cochlear duct

The space inside the cochlear partition. It is filled with fluid called endolymph. See *scala media*.

cochlea microphonics

There are first order and second order cochlear microphonics which, physically, are varying electric potentials recorded from the perilymph usually at the round window. The first order microphonics have a form very like the pressure waveform of the input sound. They are believed to have no role in hearing and originate from the organ of Corti. The second order microphonics are associated with the operation of the hair cells which also generate the nerve pulses on which auditory sensation actually depends.

cochlear nucleus

See *auditory pathways*.

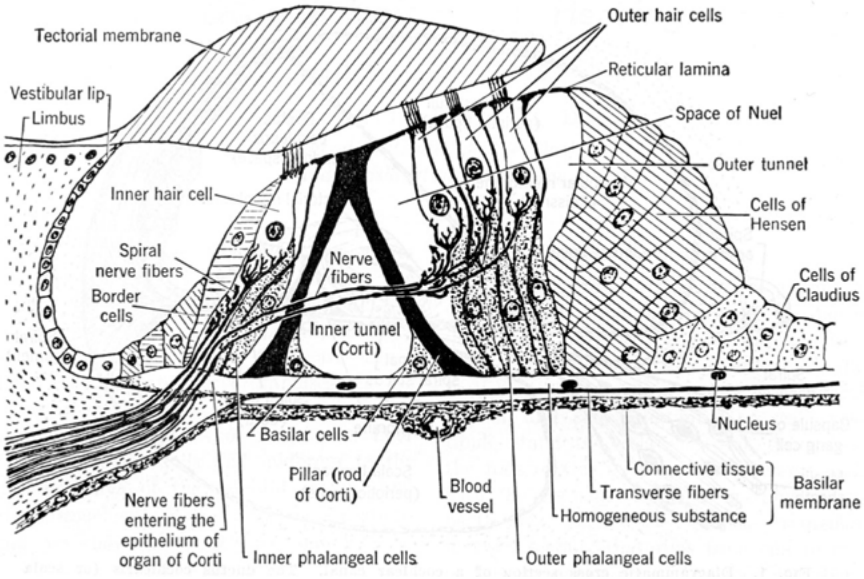

Figure 4: *Cross-section of the organ of Corti*

Cross-section of Organ of Corti p 1118 (From Handbook of Experimental Psychology, edited by S.S. Stevens.© 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

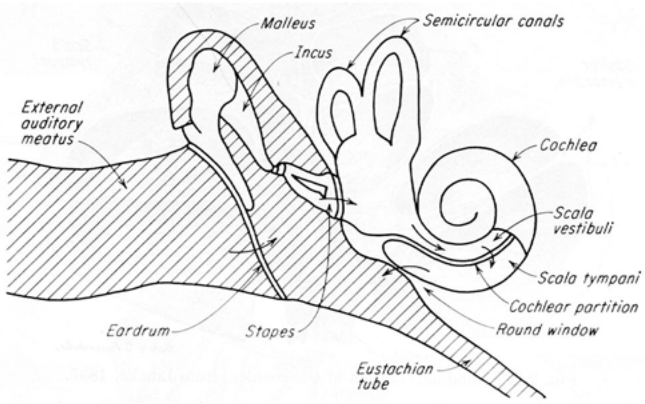

Figure 5: *Ear components*

Sectional diagram of the ear , p 11(From Experiments in Hearing, by Georg von Békésy.© 1960 McGraw-Hill Book Company. Used with permission of McGraw-Hill Book Company.)

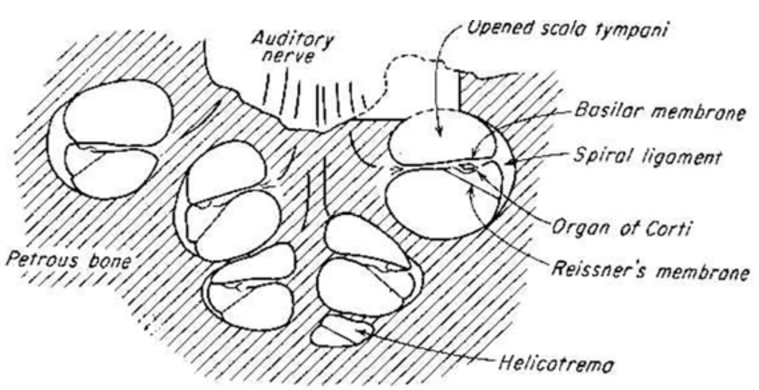

Figure 6: Cross-section of the cochlea

Preparation of human temporal bone, p 470 (From Experiments in Hearing, by Georg von Békésy. © 1960 McGraw-Hill Book Company. Used with permission of McGraw-Hill Book Company.)

cochlear partition
See *cochlea*.

cocktail party problem
Idiomatic name (derived by analogy) for the problem faced when speech is masked by other speech—the problem, in fact, faced by conversationalists at cocktail parties! Humans have unexplained abilities in accomplishing this task, though some cues (common pitch structure, signal level, directional cues, lip reading cues, semantic context, voice quality) all undoubtedly play a part. Understanding someone at a cocktail party, and understanding a recording taken of the same conversation are quite different tasks, the latter being much more difficult.

coding
The process of mapping one set of symbols onto another. Coding may be carried out in order to obtain information security (encryption), to combat noise (by building in noise-resisting redundancy), or to maximise the rate of information transmission in a communication channel (by matching the characteristics of the source of information to the characteristics of a channel. Morse code, invented by Samuel Morse long before Shannon derived his important results on the subject, provides an excellent match between English letter frequencies and a binary channel.

compression
A term applied to the process of removing redundancy from collections of information. Such collections contain more physical data than are necessary to represent the information content (information in the sense developed by Shannon). Reducing the physical data to the absolute minimum needed to represent the information results in maximum compression. It has been estimated that the change between typical successive frames of video information can be represented by less than two bits, the bulk of the information in successive frames being the same.

conductive deafness
Deafness due to some defect in transmitting sound pressure waves from the air to the oval window. See also *nerve deafness*.

confusion matrix
A square matrix having rows corresponding to (say) stimuli and columns (say) to responses, showing, in each cell, a number which represents the probability of confusion, or number of confusions in a given test, occurring between the various stimulus choices when making responses, during a psychophysical experiment. For example, a number of words may be presented to subjects under some condition of noise or distortion and listeners responses recorded. Mistakes will be made, and the pattern of confusion can be preserved as a confusion matrix. In using such data it is vital to distinguish significant confusions from those due to chance. The conditions of such an experiment are such as to cause mistaken responses. It is only the systematic mistakes that are of interest. Mistakes that are not due to some systematic effect will be distributed among possible responses on a purely chance basis. The usual statistical techniques may be used to set levels of significance and detect confusions that represent systematic effects, thereby increasing our understanding of speech perception or production.

connotation
The associative or indirect meaning, as opposed to the direct meaning—implying the attributes while denoting the subject. Intensive as opposed to extensive indication of some set. See also *denotation*.

consonant
As opposed to vowel. A stricter dichotomy applies between related contoids and vocoids. Consonants (contoids) are distinguished from vowels (vocoids) by a higher degree of constriction in the vocal tract.

context
The surroundings or co-occurring circumstances of an entity or situation. In language, a phoneme occurs in the context of other phonemes, during utterances, and the phonetic context (among other things) determines the particular allophone that occurs. Allophones are different acoustic variations of the basic (abstract) phoneme category. At a higher level, a word occurs in the context of other words. If the word is not clearly heard, there will very likely be enough contextual redundancy to allow the word to be recognised by inference.

contour spectrogram

In a conventional spectrogram, the amount of energy at a given frequency and time is represented as grey-scale values which are not easily quantified. The apparatus may be modified by including a circuit which generates pips each time the marker output crosses one of a set of internally generated reference levels. In the final analysis these extra dots join up to form contours of energy on the output just like the contours marking the height of the ground on a map. The grey scale marking is still generated and the result is a highly readable, quantifiable energy display.

contralateral

The opposite side.

coronal section

A section through an organism orthogonal to the axis running from head to tail.

cortex

The outer layer or portion of an organ—especially the outer layer of the brain, which is newest in evolutionary terms.

Corti ganglion

See *auditory pathways*.

cps

Cycles per second. Now termed Hertz, abbreviated Hz.

creaky voice

See *glottal flap*.

critical band

If a pure tone is just masked by white noise covering the entire audio spectrum, it is found that the bandwidth of the noise may be progressively reduced, without affecting the masking, until it covers a band around the frequency of the pure tone of a certain critical width. Further reduction of the white noise band, either raising the lower frequency limit, or lowering the upper limit, causes a tone that was just masked before to become audible. This band is called the critical band. It is thought (Broadbent, for example) that it may correspond to some kind of neurological "catchment area" on the basilar membrane. The width is around 50 hz at 100 hz, and increases to 1000 hz at 10,000hz. Between 50 hz and 2,000 hz the critical band remains between 50 and 100 hz and then rises roughly as the logarithm of frequency. Components of the noise outside the critical band do not contribute to the masking of the tone. (This fact should not be allowed to confuse the issue when a tone is masked by noise outside the band (e.g., a high frequency tone may be masked by a low frequency tone with around a 40 db threshold shift).

cross-talk

Leakage of unwanted signals into a communication path from adjacent paths. An extreme case occurs at cocktail parties, as far as speech is concerned, but (for example) cross-talk between telephone lines causes problems (the term "line" here includes the telephone exchange).

CSR

Continuous Speech Recognition. Automatic Speech Recognition in which the speaker speaks naturally, rather than trying to pause between words.

cybernetics

A term originated and defined by Norbert Weiner as "the science of control and communication in animals and machines". The term has been much degraded since then (there is a book entitled "The Psycho-Cybernetics of Sex" for instance), but it is still useful, with care.

cycles per second

A term replaced by Hertz (Hz). The term is self-explanatory, allowing that a "cycle" is some repeating progression of entities, operations or values.

cyton

The body of a nerve cell as opposed to its various processes.

damage risk

In relation to hearing this refers to the likelihood that a given noise exposure will result in permanent hearing loss (NIPTS—noise induced permanent threshold shift). The subject is complicated since exposure history, noise spectrum and other characteristics all contribute. The likelihood of NIPTS is usually inferred from TTS (temporary threshold shift) data measured in the laboratory. In general, a noise that produces no TTS will produce no NIPTS, if exposure is limited to no more than eight hours per day, with a 16 hour rest period between exposures.

db

See *decibel.*

dbm

A decibel scale based on a reference value of 1 milliwatt in 600 ohms and used by broadcast engineers. The equivalent voltage across a 600 ohm register is 0.775 volts-rms. A steady tone is required for calibration. Peak program meters are based on this reference, but calibration marks are not in dbm directly. See *rms*

decibel

A dimensionless unit of power (energy) measurement, abbreviated to db, which expresses power as a ratio between the target power and some defined reference: thus, for some measured power, W, db re. Wo = 10 log (W/Wo) (taking logs to the base 10).

The constant "10" is a multiplier to convert the "bels" resulting from the application of the basic formula into decibels (1 bel equals 10 decibels). The reference (in this case Wo) must always be stated. The standard for speech and noise is approximately 10-12 watts/m2. The reference or base is very often given as a pressure, rather than a power. Thus the base for speech and noise in terms of pressure is 0.0002 dynes/cm or 2 x 10-5 nt/m2, but translating this into power (energy) requires a knowledge of the air characteristics prevailing at the time. This emphasises that any db comparisons based on pressures must be transmitted in a medium of identical acoustic impedance. Under these assumptions, power being proportional to the square of pressure, the formula becomes: db re. po = 10 log (p2/po2) = 20 log (p/po). Being logarithmic, the measure is well suited to quantifying the large range of intensities encountered in acoustic measurement.

DECtalk

The most popular and successful text-to-speech system over the last few decades, based on work by Jonathan Allen, Dennis Klatt, their colleagues and their students on MITalk at MIT. The system uses a formant synthesiser, togther with a dictionary, letter-to-sound rules and some grammatical analysis to convert ordinary text into spoken words. There are a number of look-alikes and derivatives. The work is well described in the book by Allen, Hunnicutt and Klatt: *From text to speech: the MITalk system*, published in 1987 by Cambridge University Press.

dendrite(s)

The input process(es) of a nerve cell.

denotation

The meaning of something by direct example. Extensive as opposed to intensive definition. See also *connotation*.

dental

To do with teeth. A dental fricative would involve the teeth forming part of the constriction in the vocal tract that produced the fricative noise.

dichotic

See *monotic*.

difference limen

Limen is the Latin for threshold (plural limina). The difference limen is that difference or change in the stimulus that is at the threshold of detectability (since it varies from moment to moment in a random manner, it is normally defined as the change or difference that will be detected in 50% of trials). It is abbreviated to DL and is the same as the Just Noticeable Difference, or JND. The Reiz Limen or RL is the absolute threshold of detectability, below which the stimulus is not noticed at all in 50 per cent of trials. The TL or Terminal Limen is that limit beyond which the stimulus elicits pain. The DL is not constant for a given dimension of stimulus, as might be supposed.

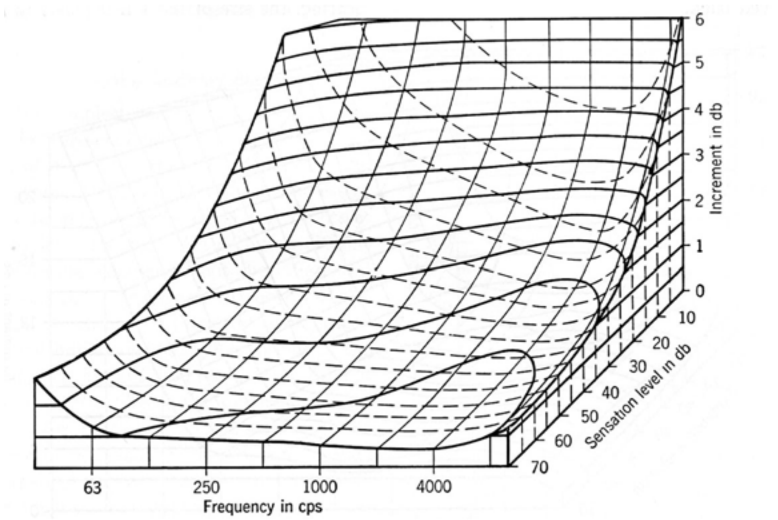

Figure 7: *Differential intensity threshold versus frequency and intensity*

Differential intensity threshold versus frequency and intensity p 999 (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

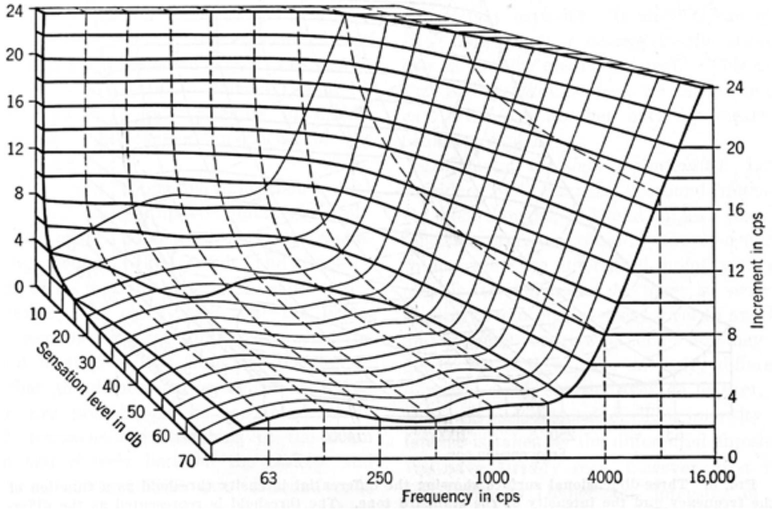

Figure 8: *Differential frequency threshold versus frequency and intensity*

Differential frequency threshold versus frequency and intensity, p 1000 (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

If we let ΔI stand for the DL at a given level of stimulus, then to a very close approximation: $\Delta I/I = K$ (a constant) (Weber's Law) where I is the intensity of the stimulus in the dimension concerned. In words "a stimulus must be increased by some fixed percentage of its current value for the difference to be noticeable". At 50 grams a 1 gram increment in a weight might just be noticed. At 100 grams, approaching 2 grams would be necessary. The following two figures show how the intensity and frequency difference thresholds vary with frequency and intensity of sounds. Note that Weber's Law is effectively incorporated into a measure of the DL based on decibels, and does not hold exactly for speech frequency and intensity. See also *method of average error, method of limits,* and *method of constant stimuli*

digram/digraph

Two letters appearing together, in order. A study of digram frequencies may be of value in automatic language processing, especially coding and compression.

dimensionless

Physical units and quantities are normally associated with dimensions of mass (M), length (L) and time (T). Thus acceleration has dimensions of L x T^{-2}. Some units, however, are dimensionless since they represent ratios of quantities of the same kinds, or mixtures of dimensions that cancel. Thus any db scale value is dimensionless— a pure number, being a ratio of two quantities of the same type.

diotic

See *monotic*.

diphone

A combination of two phones representing the instantiation of two successive speech postures. Since the interaction effects between the postures concerned are explicitly represented in diphone segments it is attractive as a basis both for synthesis and recognition on the theory that segmentation is carried out at places of minimum rate of spectral change, leading to a reliable algorithm for recognition, and simple assembly for synthesis. The success (or lack of it) of the approach reflects the validity of the underlying assumptions which are largely unstated. Thus synthesis based on assembling synthetically generated diphones, which have been carefully evaluated by relays of naive listeners, is relatively successful. Recognition has enjoyed less success, because appropriate segmentation is uncertain and difficult. Many "diphone" segments really comprise more than two segments because consonant clusters, for example, do not "obey the (unstated) rules". Whilst, in principle, around 1600 diphones could represent all the posture combinations in English, in practice, at least 3000 are required, and adding more offers much scope for improvement because coarticulation effects range over more than just two neighbouring speech sounds, so that at least 404 combinations of "tetraphones" (more than 2.5 million) might be considered necessary to account for all the phonetic level interactions, even ignoring other influences.

diphthong

A sound formed by time-sequential combination of two simple vowel articulations. Diphthongs are frequently regarded as phonemes in their own right. This is almost inevitable in view of the classical definition of phoneme, but is not too helpful for simple synthesis and recognition by machine. Consonant clusters are not regarded as different phonemes (except for affricates), so why should vowel clusters be regarded as separate phonemes (the standard reply refers to their distributional characteristics and, for the affricates, to the problem of juncture). There may be considerable modification of the component sounds compared to their "normal" realisation (reduction and shortening). Glides, which are very close to diphthongs are similarly considered phonemes in their own right, though they may be closely approximated by using segments of the related vowels, again shortened, and with diphthong-rate transitions. However, the glides also represent a much greater obstruction of the oral cavity than the related vowel sounds (which also leads to a somewhat different spectral character) so that they are properly identified as contoids (i.e. loosely speaking as "consonants"), rather than diphthong combinations. See triphthong.

diplacusis

A disparity in pitch perception between the two ears. One common way of inducing diplacusis is by using a loud tone to fatigue one ear. Judgements of pitch in the region of the fatiguing tone may then differ by as much as an octave between the two ears. The condition can exist to some degree, in many subjects, as a natural state. It may explain why unaccompanied folk singers conventionally stuff a finger in one ear whilst singing.

distinctive features

Distinctive features are defined for speech sounds as binary distinctions between polar extremes of a quality, or between presence versus absence of some attribute of the sound. It is an approach to speech segment description based upon Daniel Jones' idea of "minimal distinctions"—any lesser distinction between two sounds being incapable of distinguishing two words clearly. Distinctive features are thus closely related to phonemes and subject to the same limitations. Phonemes may be considered as concurrent bundles of distinctive features. Problems arise with consonants. (*Preliminaries to Speech Analysis: the distinctive features and their correlates*, by Jakobson, Fant and Halle, MIT Press 1969, is the original reference to this approach).

DL

See *difference limen*.

DTW

See *Dynamic Time Warping*.

Dynamic Time Warping

A process for normalising the time relationship between an unknown speech waveform and a reference waveform, carried out by reference to frequency domain data. If the corresponding parts of the varying speech signal for the reference and the unknown can be mapped onto each other, the problem of deciding whether they represent the same speech sounds is made much easier. See also *time normalisation*.

dyne

A unit of force, being the force required to give one gram of matter (a mass of 1 gram, that is) an acceleration of 1 cm/sec2. It is now obsolete, having been replaced by the newton in the SI system of measurement. One newton gives 1 kg mass acceleration of 1 m/sec2 and is thus 100,000 dynes.

dynes/cm2

A measure of pressure. 1 dyne/square centimeter = 0.1 nt/square meter = 0.1 Pa/square meter = 1 microbar = 0.011 kg/square meter. Because of the wide range of pressures involved in acoustics, a logarithmic function of power ratios is used—the decibel scale. Because this scale is based on power (energy) ratios, there must be an assumption of constant impedance in the transmission medium, when expressing pressure ratios, and the pressures must be squared, or the logarithm of the pressure ratio multiplied by 2, to convert to an energy ratio. In air at 20 degrees Celsius and standard pressure, 1 dyne/square centimetre is equivalent to a power or intensity of: 2.41 x 10-4 watts/square metre (i.e. about 1/4 of a milliwatt) See *acoustic intensity*, and *decibel.*

E/E's

Experimenter/experimenters.

ear drum

The membrane dividing the outer ear (external auditory meatus) from the middle ear—the air-filled cavity containing the ossicles. The membrane has a shallow conical shape with the "handle" of the malleus ("hammer" or first ossicle) lying radially on the upper vertical radius, being hinged at the top edge. The bottom edge of the membrane has a more or less pronounced fold, which frees the lower edge and allows the ear drum to rotate the malleus about its hinge. As a result, the ear-drum acts as a piston in the sound transmission chain. Impedance matching with the air is achieved by the mechnical arrangements created by the ossicles and also by the construction of the basilar membrane/organ of Corti/tectorial membrane combination. Figure 6 provides a diagrammatic view of the ear mechanisms. See *cochlea.*

effectors

Organs (devices) for operating on the environment of an organism (or machine).

efferent

Conducting outwards, towards the periphery. Thus efferent nerves are those which carry impulses from the central nervous system to the effectors (muscles and glands).

enclitic

See *proclitic*.

endolymph

Viscous fluid filling the cochlear partition (or duct). See *cochlea*

envelope

Roughly, the "shape" which contains components forming an object of interest. In mathematical terms, it is similar to the convex hull, though will usually involve concave portions as well. The real criterion is the amount of detail that the shape reveals. Thus the envelope of a narrow band spectral section would approximate a wide band spectral section. The envelope of a speech time-waveform would approximate the function obtained by joining all the positive peaks together and all the negative peaks, smoothing out any significant kinks. The form of an envelope function depends quite a lot upon the smoothing function applied to the original function.

ergonomics

From the Greek "ergon" work—literally ergonomics means "the study of work", but is actually the British term for human factors engineering, which involves a study of the working conditions most suited to the many tasks an operator is asked to perform.

esophageal speech

see *oesophageal speech.*

eustachian tube

A narrow canal which connects the middle ear to the throat. It is normally closed in humans, except when swallowing or blowing. In some creatures, notably the frog, a very wide short eustachian tube acts to make the ear drum act like a noise-cancelling microphone for self-generated sounds, with obvious benefits.

external auditory meatus
See *auditory meatus*

FH2
The frequency of a filter used in a Parametric Artificial Talker to simulate the spectrum of sibilant sounds (fricatives -- such as those at the beginning of “sun” and “shun”), which are produced by forcing air through a narrow constriction located towards the front of the oral cavity.

Fx/Fo
The glottal vibration frequency or “larynx” frequency in speech, especially in synthetic speech. Fo is strictly the fundamental frequency of a repetitive time-waveform. The glottal vibration frequency is only quasi-periodic, but it is convenient to treat the glottal waveform as instantaneously periodic (sic) and talk about the fundamental frequency when analysing the frequency spectrum of speech. See *fundamental frequency*

F1/F2/F3
Formant 1/formant 2/formant 3. As speech synthesiser parameters these represent the frequencies of the resonant peaks in the output spectrum. There are higher formants (F4 ... Fn) but these are less important in making primary distinctions between speech sounds. See *formant* and *articulatory synthesis*.

falsetto
A mode of voicing (principally associated with male speech) in which an unnaturally high Fo is produced by setting up an abnormal mode of glottal vibration.

Fast Fourier Transform
A method devised by Cooley & Tukey (*Mathematics of Computing* 19, 297- 301, April 1965) for computing the coefficients of the discrete Fourier transform (the digital computer equivalent of the Fourier transform). The method is very much faster than previous algorithms and a considerable literature has developed on this special topic because of its importance in signal analysis, and the widespread use of computers in laboratories around the world. See also *Linear Predictive Coding (LPC)*.

FFT
see *Fast Fourier Transform*.

filter
A transmission device of limited bandwidth.

foot
Unit division of speech according to the occurrence of stressed syllables (“beats”), rather as music is divided into bars. Each foot begins with a syllable bearing primary stress, and ends just before the following primary stress.

formant
A formant is a significant peak in the spectral envelope of the frequency spectrum characterising speech. In general, it varies as a function of frequency against time. The lowest three formants are considered adequate for good intelligibility, though in synthetic speech using a so-called “formant synthesiser” additional higher poles are required to compensate for the missing higher formants and the radiation impedance from the mouth. Formants appear as dark bands in a broad-band spectrogram. See *articulatory synthesis* and *sound spectrograph*.

formant synthesiser
A “terminal analogue” type of synthesiser which models speech in the acoustic domain by using filters to simulate the resonant behaviour of the vocal and nasal passages. Pulsed or random energy is fed into the input to simulate the glottal pulses and/or aspiration, and additional random energy, shaped by filters is added to simulate fricative energy. The intensities are controlled directly and the energy balances are not well represented. Moreover, only the lowest three or four formants are properly represented and varied. The resulting voice quality is less than natural, even discounting the difficulties of representing dynamic variation, timing and intonation accurately. This “terminal analogue” is an input-output representation of the vocal tract behaviour rather than a model of its distributed acoustic properties. It is contrasted with a transmission-line or acoustic tube model which directly models

the acoustic properties of the vocal apparatus as a collection of tubes and energy balances. The author was involved in the development of a synthesis system of the latter type. See *articulatory synthesis*.

formant tracking
The process of determining the frequency position of formant peaks automatically, especially the first three formant peaks. It is not as easy as it might seem. Decision criteria to distinguish significant peaks from insignificant peaks have to be set, and will be somewhat arbitrary; formants may come so close together that they form a single peak; and in some sounds the formant 1 peak may be considerably reduced in amplitude, or split by a nasal anti-resonance. Furthermore, formant peak frequency values may change very quickly in just those places where accurate determination is most important (e.g. the transitions associated with a stop sound), yet one strategy for eliminating noise from a formant tracker output is to apply continuity criteria to the values obtained. Clearly it would be useful to have a mechanism capable of making a binary distinction between rapid and slow movement, adjusting the application of the continuity criteria appropriately. The problem is one of many. It is typical of the problems encountered in the automatic analysis of speech.

frequency analysis
An analysis of a time varying signal into its individual frequency components. Also known as Fourier analysis.

frequency spectrum
The total range of frequencies needed to contain all components of the Fourier analysis of a sound. The frequency domain description of a phenomenon is, loosely, called its (frequency) spectrum. See also *harmonics*.

Fourier analysis
The decomposition of a complex waveform into sinusoidal components, and determination of the coefficients (amplitudes) and phase of the various terms. This gives a Fourier series. Any physically realisable repetitive waveform can be represented to arbitrary accuracy. In dealing with non-repetitive waveforms there are problems. An impulse produces a uniform spectral density function in place of a Fourier series. For non-transient waveforms, a segment of the waveform is excised, and treated as if repeated indefinitely. This arbitrarily selects the collection of sinusoids that underlies the transform and, in particular, implies a non-existent fundamental. In speech, which, strictly speaking, is non-repetitive, an excellent compromise is achieved by choosing the time-interval between successive glottal pulses as the unit for decomposition in this repeated segment fashion—making the analysis so-called "pitch-synchronous". Spectrograms computed on a pitch-synchronous basis show considerably more coherence than those computed on the basis of arbitrary segments. Of course, the method should be called glottal-pulse synchronous. See *pitch* and *harmonics*.

Fourier series
That which is produced by *Fourier analysis*.

Fourier transform
See *Fourier analysis*.

frequency
The repetition rate of a repeating event. The reciprocal of the time-interval between successive repetitions of the event, especially successive repetitions of a cyclic waveform.

frequency analysis
See *Fourier analysis*.

frequency domain
A description of a system or phenomenon based on frequency and phase measures. See *time domain*.

frequency spectrum
The range of frequencies allocated, or considered, or involved in dealing with frequency-related phenomena.

fricative
A speech sound produced by forcing air through a narrow constriction, thereby generating noise due to air turbulence which is characteristic of the constriction—so-called friction noise, and is somewhat shaped by any

coupled cavities. If the vocal folds are maintained in an open position during speech articulation, so that they do not vibrate and there is no voicing, then with constriction higher up in the vocal tract (say between tongue and palate) an unvoiced fricative results. The acoustic correlates of an unvoiced fricative are: the offset and onset of voicing; early disappearance and late reappearance of F1; formant transitions; and a spectral energy distribution appropriate to the particular place of articulation. If the vocal folds vibrate, then a voiced fricative results. The friction energy of a voiced fricative is usually somewhat modulated in amplitude at the voicing frequency. During the constrictive part of a voiced fricative there is a drop in the pitch frequency (one form of micro-intonation) due to the supraglottal pressure increase caused by the vocal tract constriction. See also *voiced stop*.

fundamental frequency

The lowest component frequency in the analysis of a periodic waveform. See *Fourier analysis*.

glottal flap

Excitation of the vocal tract by what are effectively isolated glottal pulses. This happens typically at the ends of utterances by some male speakers, as the voicing frequency (pitch) is allowed to fall dramatically. It is similar to creaky voice, which occurs at the beginning as well as the end of utterances for many speakers of educated Southern British English ("RP" from "Received Pronunciation", the accent that was once expected of professional radio announcers and journalists in Britain). In both, the glottal rate can reach abnormally low values, and the pulses are then so well separated that they are easily seen as separate in a broad band spectrogram.

glottal pulse

The airflow pulse resulting from one cycle, opening and closing, of the glottis during voiced speech. One cycle of the glottal waveform

glottal rate/frequency

The frequency of vibration of the vocal folds that surround the glottis, creating "lips" that can open and close to control the air flow. Normally the repetition rate is not constant, even for two successive cycles.

glottal waveform

The volume velocity waveform of air through the vocal folds (glottis) during voiced speech. In normal speech, it approximates a triangular waveform at normal voicing effort, with soft opening and sharp closure. See *harmonics*.

glottis

The opening that varies from nothing, through a slit of increasing width, up to a wide open triangle formed between the vocal folds. When the tension and position of the vocal folds is suitably adjusted, and air is pressure applied from the lungs below, fairly regular puffs of air break through the lips of the vocal folds and provide so-called "voiced" excitation of the resonant cavities of the vocal apparatus. The vocal folds are located in the larynx.

hair cells

Cells lying in the organ of Corti which provide mechanical-to-electrical conversion of sound vibrations as rendered at the tectorial membrane, and hence generate electrical activity in the auditory neurons (from which arise the fibres of the VIIIth cranial nerve, which also innervates the vestibular apparatus). Each cell has small hair-like processes which are mechanically stimulated by lateral movements of the tectorial membrane, with respect to the organ of Corti. These movements result from the displacement of the basilar membrane caused by pressure waves transmitted into the cochlea. It is of interest that the neurons of the auditory nerve, unlike most others, do not regenerate following injury to the processes, but generally die. The hair cells innervated by such a neuron then also die. It was shown in 1988 that the equivalent cells in chickens regenerate after damage (*Science*, June 1988, Corwin, University of Hawaii and Cotanche, Boston University). This holds out hope that, one day, it may be possible to help or even cure humans with noise-induced hearing loss. See *cochlea.*

half octave band spectrum

See *octave band spectrum*.

hard palate

The roof of the mouth (oral) cavity, which divides it from the nasal cavity. It lies between the velum and the teeth. See *alveolar ridge*.

harmonics

When a note is played on a musical instrument, a sound is produced having a certain pitch, with overtones. The pitch is closely related to the fundamental frequency of vibration of the mechanism producing the note, while the overtones, comprising some selection of frequencies that are integer multiples of the fundamental, give the instrument its characteristic timbre. Frequencies falling at integer multiples of the fundamental frequency are called harmonics and a Fourier analysis of a repetitive waveform effectively decomposes a complex waveform into the fundamental and its harmonics. A triangular waveform (symmetrical linear rise and fall during the first and second half cycles respectively) contains all odd harmonics. So does a square wave (being a differentiated triangular wave, this is a logical consequence). A triangular wave is a special case. Any waveform that has a rise and a fall in two linear sections comprising a full cycle will contain all harmonics except those that exactly divide both rise and fall periods an integer number of times (not necessarily the same for the two sections). The glottal waveform is slightly assymetric and therefore contains most harmonics in the speech frequency range. Missing or reduced harmonics can affect the output spectrum dramatically when convoluted with the filter function of the vocal tract. The characteristic sound quality of a church bell results from the rather strange distribution of frequencies it produces which are by no means all harmonically related. A tubular bell, which rings with a normal set of harmonics produces a much less interesting sound than a church bell.

Haskins Laboratory

One of the two or three most important speech research laboratories in the world. Much of the original work on elucidating speech cues was carried out there, and some of the first serious synthetic speech was produced there in the 1950s using a spectrogram playback apparatus known as Pattern Playback. Frank Cooper, Pierre Delattre, Alvin Liberman, Leigh Lisker and many others laid the ground-work for speech synthesis by rules. Their current research work is now available on their website, replacing their well-known annual reports. They are now located in New Haven, CT, as part of Yale University.

hearing loss

A shift in hearing sensation level and thus defined as: HL = 10 log (I/Io) db where I is the intensity (sound energy) at the threshold of hearing for the patient, and Io is the intensity at the threshold of hearing for normal persons—a statistical average over normal subjects). See also *damage risk*.

helicotrema

The sole connection between the *scala vestibuli* and the *scala tympani*, the two spaces either side of the cochlear partition. The helicotrema allows volume displacement of the fluid in the scala vestibuli by the oval window movements caused by the stapes to be transmitted into the scala tympani, ultimately causing volume displacement of the round window. This streaming of the fluid, induced by pressure waves on the ear drum, produces maximum pressure differences across the cochlear partition, and hence distortion of the basilar membrane, at places corresponding to frequency of excitation (for sounds in the normal hearing range). As a result, nerve pulses are generated in different fibres corresponding to different frequencies. The pulses themselves are also grouped to reflect the frequency and phase of the sound. The highest frequencies produce displacements of the basilar membrane near the windows, the lowest near the helicotrema, the range being about 20 khz down to 20 hz. See *cochlea, basilar membrane, volley theory*.

Hensen's cells

Cells forming part of the organ of Corti and capable of generating steady voltage potentials.

Hertz

The unit for measuring repetition rate of a repeating event, especially a regularly cyclic event. Abbreviated to Hz.

Hidden Markov Model

A technique for recognising speech based on a specialised state machine in which the states represent the varying spectrum of the speech signal, classified into discrete categories, and linked by transitions. Different segments of speech drive the HMM through different paths. The path allows the segment or succession of segments and therefore the speech input to be identified. There is a vast literature on the topic, as the algorithms for creating, searching, backtracking and so-on within the network, are very important to economy and success. See also *segment synchronisation*

high back vowel
See *vowel.*

high front vowel
See *vowel.*

HMM
See *Hidden Markov Model.*

Hz.
An abbreviation for Hertz.

IEEE-ASSP
Institute of Electrical & Electronics Engineers Transactions on Acoustics, Speech and Signal Processing. An important source of current research with an emphasis on electronics and signal processing. Published by the Institute of Electrical & Electronics Engineers.

IJMMS/IJHCS
International Journal of Man-Machine Studies, now renamed the *International Journal of Human-Computer Studies*. An important source of current research with emphasis on all aspects of systems involving Human-Computer Interaction (HCI, also known as Computer-Human Interaction—CHI). Currently published by Elsevier

incus
See *ossicles*.

inferior colliculus
See *auditory pathways*.

information
Information is the subject of Claude Shannon's original theories on the transmission of information (*Bell System Technical Journal* volume 27 1948: A mathematical theory of communication, pages 379–423 and 623–656). Information is measured as the logarithm of the probability of a message. In the case of logarithms to the base "2", the measure becomes a bit for binary digit. Only when there is no noise, and the possibilities of a "1" or "0" are equally likely, can a physical "1" or "0" (a "binit" to distinguish it from the information measure) convey one bit of information. The problems of coding to deal with noise, or unequal message probabilities, are the province of information theory and coding theory. See *redundancy*.

information theory
See *information* and *redundancy*.

inner ear
The cochlea and all it contains.

intelligibility
A sound may or may not be heard. If it is heard, it is audible. To be intelligible, it must be a speech signal, and the listener must recognise the nonsense syllable, word, sentence, or whatever that was sent to him. The threshold of intelligibility is some 10 to 15 decibels higher than the threshold of audibility for the same speech under the same noise conditions (Hawkins & Stevens 1950). Tests of intelligibility are used to evaluate communication systems and situations. See *articulation index*.

intensity
See *acoustic intensity*

intensity spectrum level
Is defined as the acoustic intensity per Hertz for the noise in question. If the band of frequency containing the noise is ΔF hz wide, then: ISL = 10 log(I/(Io.ΔF)) db re. Io = IL - 10 log ΔF
where IL is the sound intensity level. See also *pressure spectrum level, acoustic intensity, sound intensity level, sound pressure level* and *decibel.*

interval scale

See *scales of measurement*.

intonation

Often referred to as the "tune" of an utterance. The intonation pattern of an utterance is the time-pattern of pitch-variation during an utterance and serves several purposes at different levels. At the segmental level, so-called "micro-intonation" provides cues to constrictive postures (contoids) of the vocal apparatus because the voice pitch (glottal frequency) rises and falls as the pressure difference across the glottis varies as a result of the changes in supraglottal airflow caused by the varying constrictions. At the prosodic (suprasegmental) level, the intonation pattern gives clues to syllable, word, and sentence structure. At the semantic level intonation affects meaning. It should not be thought that intonation achieves its effect in isolation, however, even though it would have a considerable effect under such conditions. Of considerable importance is the precise relationship between the pitch movements and levels that make up the intonation pattern, and the segmental features, rhythm cues, and perceived loudness effects that run in parallel. In tone languages (such as Chinese), intonation also directly affects the meanings of words, since words with different meanings may differ only in the tone applied. See *stress*, *prosody*, *salience,* and *prominence*.

IPO

The Institut voor Perceptie Onderzoek, part of the Technical University of Eindhoven in the Netherlands. The IPO works on speech perception research, and computer-human interaction amongst other topics. They have carried out seminal work on English intonation.

ipsilateral

The same side.

isochrony/isochronicity

The theory of isochrony states that in British English different rhythmic units in speech (feet—Abercrombie/ Halliday, rhythm units—Jassem) have a tendency to be of equal duration regardless of the number of segments or syllables they contain. Studies have shown that the ratio between the duration of the longest rhythmic unit to that of the shortest is as much as 6 or 7 to 1. Therefore American linguists (e.g. Lehiste) consider that most of the perceived isochrony effect is just that—a perception on the part of the listener, rather than an objective tendency towards equal duration. A statistical study by the present author, with Jassem and Witten, showed that the tendency towards isochrony was one of the only three independent factors determining speech rhythm. In setting the duration of segments, when modelling speech rhythm, the identity of segment accounted for about 45% of the variance in duration; whether the rhythmic unit was marked (tonic or final) accounted for about 15% of the variance in duration; and the segment durations then needed to be corrected by an inverse linear regression based on the number of segments in the rhythmic unit, accounting for about 10% of the variance in duration. The effect was almost non-existent for proclitic syllables (anacruses in Jassem's terminology), and Jassem's theory therefore excludes anacruses in estimating rhythmic unit duration for purposes of testing and elaborating the theory. See salient, foot, rhythmic units and anacrusis.

IWR

Isolated Word Recognition. See *automatic speech recognition*.

JASA

The *Journal of the Acoustical Society of America*. A basic source of important work in speech, psycho-acoustics, and other topics. Published by the American Physical Society and extremely well indexed.

JSHR

The *Journal of Speech and Hearing Research*. Published by the American Speech and Hearing Association.

JVLVB

The *Journal of Verbal Learning and Verbal Behaviour*. Published by Academic Press.

juncture

The difference between the utterances "night rate" and "nitrate" is one of *juncture*—where does the word boundary fall. Juncture is important. Affricates in spoken English might be regarded simply as two successive speech postures in a row: but, for reasons of juncture and for reasons of distributional characteristics, they are treated as distinct single phonemes. For example, “my chip” has the word boundary before the combined sound of “t” followed by “sh”, whilst in “might ship”, the juncture is in the middle. The combined “t-sh” (IPA /ʧ/) sound is one of the English affricates. Although the dynamic acoustic elements are spectrally quite similar to the related simpler phonemes, the timing of the elements is significantly different, and signals the position of the word boundary.

Kay Sonagraph/Kay Elemetrics Sonagraph
See *sound spectrograph.*

kinaesthetic sense
A sense of the spatial relation and movement of its own frame possessed by a conscious entity. See *proprioception*.

labial
To do with a lip, lip-like part, or labia (singular labium).

L & S
Language & Speech. Another key journal. Holme’s early work on speech-synthesis-by-rules was published in this journal. Published by Robert Draper, Teddington, UK.

laryngograph
A device for determining the voicing frequency (glottal rate/frequency) when a *live* talker is producing speech and can be instrumented. Direct measurement of the frequency of vibration of the vocal folds may be determined on a pulse to pulse basis in a variety of ways involving electrical impedance, light beam modulation and so on. The original laryngograph worked by placing non-invasive electrodes either side of the speaker’s larynx (on the neck) and measuring the high frequency impedance (impedance to a signal at approximately 1Mhz). Peter Ladefoged wore “Nehru” collars to hide his resulting scars! Although the change in impedance as the vocal folds open and close is quite small, it is possible to pick up the varying contact of the folds, and convert successive intervals to a display of glottal pulses. Ideally some decision criteria as to presence versus absence of voicing should be built in to suppress indications during periods deemed voiceless. Setting up an adequate test of voicing *versus* no voicing is not as easy as it might seem.

larynx
A cartiligenous upper chamber of the trachea (or wind-pipe) that may be felt from outside the throat as the Adam’s apple. The larynx contains the glottis which is bounded by the vocal folds. Enlargement of the larynx, and hence lengthening of the vocal folds, is a secondary sex characteristic for male humans, and is responsible for the voice “breaking” at puberty, and for the lower pitch of the usual mature male human voices.

lateral
To or toward the side: away from the mid-line. Thus a lateral consonant is produced by constriction allowing airflow either side of the tongue.

lateral lemniscate and nucleus
See *auditory pathways*.

limen
Threshold. An absolute threshold marks the boundary between what is just below the level of sensation and what is just above. A differential threshold marks the boundary between a difference in sensation which is just below detection, and a difference which is just detectable. The threshold of feeling or sensation marks the boundary below which feeling or sensation does not occur and above which it does. Thresholds represent statistical results from defined groups of subjects (usually subjects with normal sensory abilities). Individual differences are important and any particular individual may differ significantly from the statistical averages. Figures 7, 8, and 9 illustrate absolutes thresholds of hearing, and differential frequency thresholds at various stimulus frequencies. See also *difference limen, method of average error, method of limits,* and *method of constant stimuli*

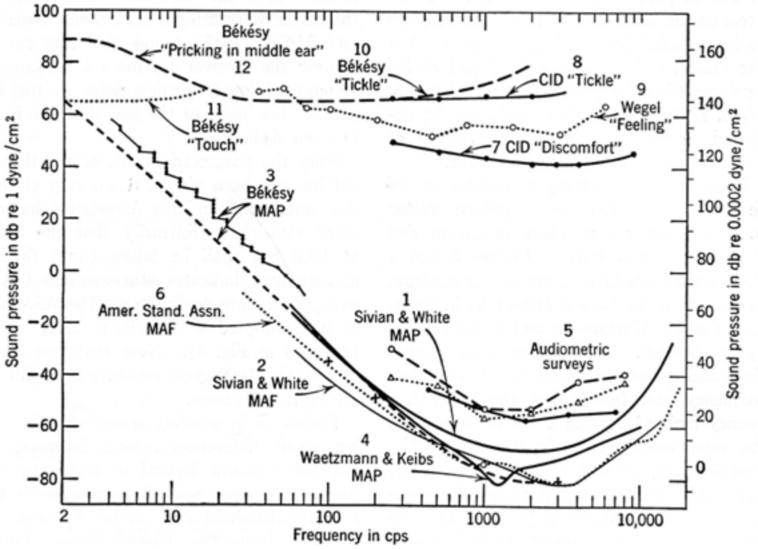

Figure 9: *Threshold of audibility and of feeling (various sources)*

Page 995,(From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

linear predictive coding

A method of representing the sampled speech speech waveform based on the coefficients of a linear polynomial, abbreviated as LPC. For analysis, the *n* polynomial coefficients are adjusted to predict the next sample value, based on *n* previous samples (hence the name). The varying values of the coefficients can be used to represent the waveform at a lower information rate than would be required to represent the time waveform directly. Thus LPC is a form of compression of the speech waveform (in information rate terms). The technique has become the method of choice for those involved in the analysis, and compression of speech waveforms. An excellent summary of the methods for computing the coefficients, and the problems, appears in Witten, I.H. *Principles of Computer Speech*, Academic Press 1982).

lingual

To do with the tongue.

lip rounding

The formation of a more or less protruding cylindrical passage with the lips.

localisation

The name given to the subjective identification of a point in auditory space as being the source of some sound stimulus.

loudness

The subjective aural quality correlated with acoustic intensity. It is not the same thing as volume. In speech, duration and pitch movement have more effect in making syllables stand out than does simple acoustic intensity. Figures 10a and 10b below illustrate some of the complexities of subjective versus objective scales of physical phenomena. The apparent loudness of a tone varies with frequency as well as intensity, and is not linear with either. The scale of sones is a subjective ratio scale (any tone with a loudness of 2 sones will sound, subjectively, twice as loud as any tone of 1 sone). However, a tone having a loudness level of 40 phons will not sound twice as loud as one of 20 phons, but two tones with the same loudness level (say n phons) will sound equally aloud. A 1000 herz tone with an acoustic intensity of 40 db is defined as having a loudness of 1 sone. Are you sufficiently confused? Try investigating colour perception! The bottom line is that perception is relative and non-linear compared to our instrumental approaches to measurement. See also *sones, stress, prominence,* and *salience*.

low back vowel

See *vowel.*

low front vowel

See *vowel.*

LPC

See *Linear Predictive Coding.*

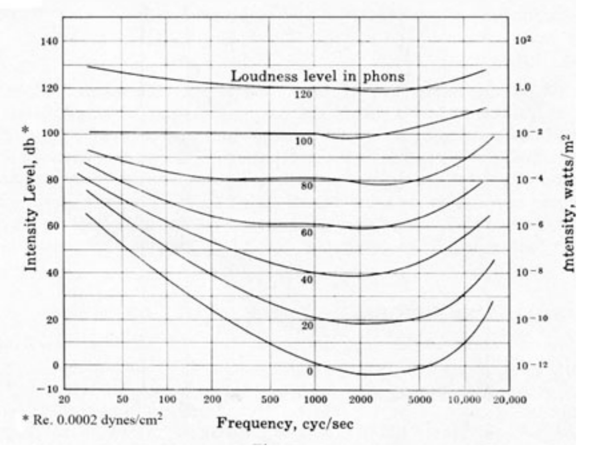

Figure 10a: *Loudness in phons plotted against frequency intensity in db re. 0.0002 dynes/cm² and Watts⁄cm²*

Page 141,(From Theory and problems of acoustics by William W. Seto. © 1971 McGraw-Hill Book Company. Used with permission of McGraw-Hill Book Company.)

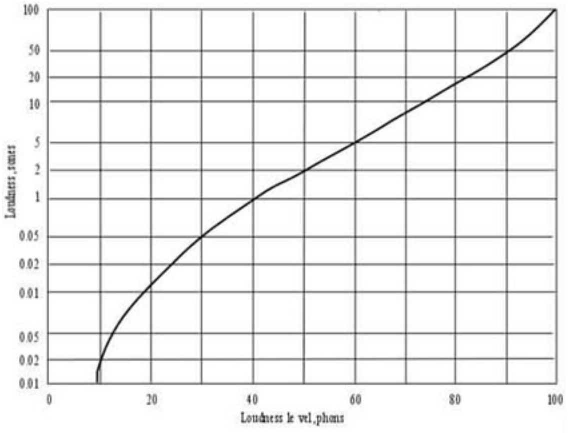

Figure 10b: *Relationship between sones (logarithmic scale) and phons (linear scale)*

Note: At 1000 Hz. a loudness of 1 sone is equivalent to a loudness level of 40 phons)

lumen
Interior space.

malleus
See *ossicles*.

masking
Masking is the process by which one sound hides the occurrence of another. The masking effect of one sound on another is measured as the increase in threshold (in decibels) for the stimulus sound in the presence of the masking sound, compared to the stimulus threshold when the stimulus is presented alone. For speech, masking effect is expressed in terms of the decreased intelligibility of the speech in the presence of the masking sound, since it is this rather than its audibility, that interests us in the case of speech. The main facts about masking are: (a) masking tends to be greater for tones close in frequency than tones widely separated; (b) low frequency tones mask high frequency tones fairly effectively but not vice versa; (c) the rate at which masking increases with the intensity of a masking tone depends on the frequencies of the tones; (d) when the tones are applied to different ears, the small amount of masking that occurs is mainly due to trans-cranial conduction (50 db attenuation) across the head; (e) speech is the most effective masking noise for speech; (f) with sufficiently intense high-frequency noise, sub-

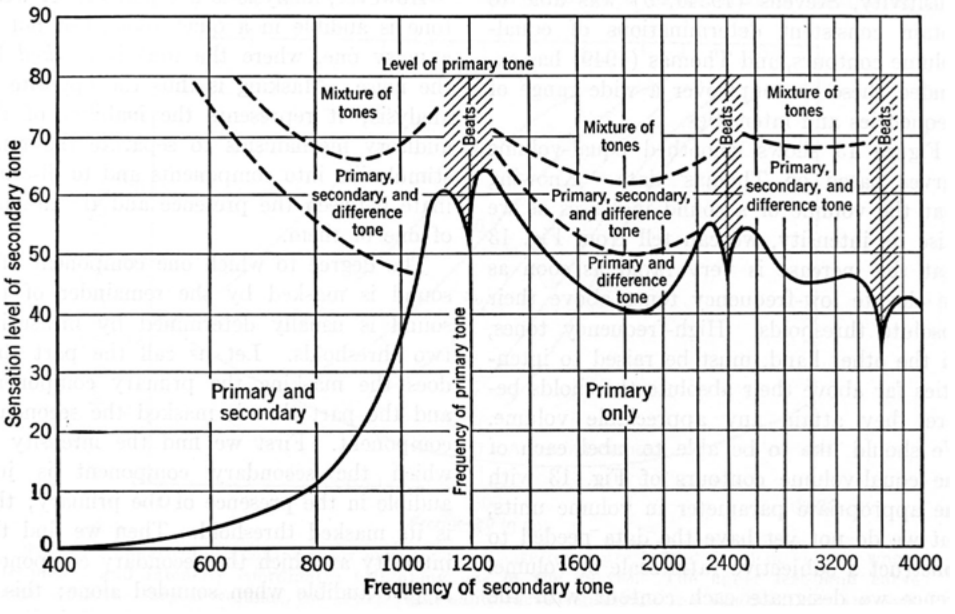

Figure 11: *Sensations produced by a two-component tone*

(Page 1006, from Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

harmonics are produced by the ear mechanism which can cause (anomalous) “remote masking”; (g) the effect of a masking noise may be displaced in time by some few tens of milliseconds, both forwards and backwards.

The complex relationships involved in masking are well illustrated by the following figure which shows the various sensations produced by the simultaneous presentation of two tones to one ear (*monotic* presentation).

maximum tolerable level

See *threshold of pain*.

meatus

See *auditory meatus*.

medial

In or towards the middle.

medial geniculate

See *auditory pathways*.

medulla

The brain stem. Controls basic autonomic functions (respiration, heart- rate) and includes structures (the reticular formation, not to be confused with the reticular membrane) which directs incoming sensory information.

mel

A unit of measurement on the subjective scale of pitch. 1000 mels is defined as the (subjective) pitch of a tone which is 40 db above the threshold of hearing at 1000 hz i.e. the pitch of a 1000 hz tone at a sensation level of 40 db is 1000 mels. Since the scale is constructed by asking for judgements of pitch ratio (half and twice the pitch of a reference are fairly readily judged) the mel scale is a ratio scale. The following plots of pitch versus frequency (using both linear and logartithmic scales) shows, the subjective pitch is not linear with frequency, nor even (as might be expected) with the logarithm of frequency. See *scales of measurement*.

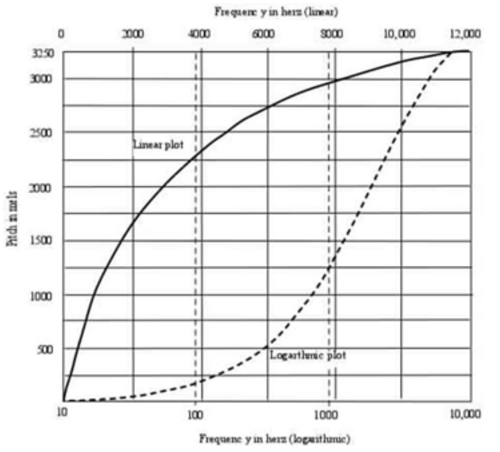

Figure 12: *Pitch in mels (subjective) plotted against both linear and logarithmic frequency (objective) scales*

Figure 13: Pitch *versus* Frequency

membranous labyrinth

This membranous structure is contained within, but distinct from, the bony labyrinth. It includes the semicircular canals (concerned with balance) and the cochlea (concerned with hearing). The interconnected membranous labyrinth is attached at various places to the bony labyrinth, and is filled with a viscous fluid known as endolymph.

meta-language

A higher level of language. A meta-language implies, of necessity, a language below it. In order to talk about language one uses a meta-language.

method of average error

A psychophysical method for determining Difference Limens (thresholds) or DL’s. Also called the adjustment method. In a psycho-physical experiment, subjects are asked to adjust a comparison stimulus with respect to some standard stimulus until the two appear equal. There will be some average error in setting which, by normal statistical methods, may be computed. It provides an estimate of the difference limen (the Just Noticeable Difference, or

JND, between values at various levels for the stimulus concerned). See *difference limen, limen, method of constant stimuli,* and *method of limits*.

method of constant stimuli

A psychophysical method of determining limens (thresholds). A set of stimuli is chosen and repeatedly presented in random order (hence "constant stimuli" really means constant set of stimuli). Some stimuli will rarely be noticed, some almost always. The value for which the probability of being noticed is 50 percent is taken as the absolute threshold, or RL (*Riez Limen*) for the stimulus concerned. The following two figures show plots of absolute thresholds and of differential thresholds for sounds. See *method of average error* and *method of limits*.

method of limits

A psychophysical method of determining limens (thresholds). For the absolute threshold the subject (S) is asked to respond on successive series of sequential stimulus intensities crossing the threshold region from above and below. A doubtful response is interpreted as a change in sign from a previous positive or negative response. Each series of stimulus intensities covers different range around the threshold and/or is of varying length, to avoid recognisable patterning. By alternating the direction, errors of anticipation or habituation are (hopefully) minimised. The results are averaged to give the required threshold. To obtain the difference limen, two stimuli are presented on each trial, one being a standard and the other a varying stimulus. Subjects are asked for three categories of response as before, with ascending and descending series of comparison stimuli, again avoiding strict patterning (otherwise a subject may count, for example). In the neighbourhood where subjective equality is obtained, subjects will give "doubtful" judgements. In each series the doubtful judgements defining the largest band of uncertainty are taken as the positive and negative difference limina (thresholds), even though the band also contains + and - judgements. Averaging over many series gives the final results. See method of constant stimuli and method of average error.

microbar

A measure of pressure. One millionth of a bar. See also *dynes/cm^2*.

micro-intonation

Perturbations of voicing frequency during speech caused by the changing pressure applied across the glottis resulting from constriction of the supraglottal airflow. For example, when a voiceless stop is articulated, the voicing frequency falls as the closure occurs, because the pressure across the glottis (vocal folds) falls due to the build up in supraglottal pressure cause by the constriction, and then vibration ceases. When the constriction is released, and voicing begins again, the voicing frequency initially starts out at an even higher level than it would have been without the stop articulation. It then falls down to meet the continuation of the smooth track extrapolated from the time prior to stop articulation. Such micro-intonation, which may involve as much as an octave shift in frequency, can give valuable cues to consonant identity. Its presence or absence in speech-synthesised-by-rules can affect both naturalness and intelligibility.

middle ear

That bony space between the outer ear, or auditory meatus, and the inner ear, or cochlea. Into it face the ear drum (leading from the outer ear canal), the oval and round windows (leading into the cochlear duct), and the eustachian tube (which connects with the throat). Thus the middle ear contains the ossicles and is connected to the throat by the eustachian tube. The tiny muscles attached to the ossicles within this space are contained in bony shells to prevent them from being set into vibration by the air-vibrations. Such induced vibrations would cause the generation of subharmonic frequencies of the input frequencies. This one of many wonders of the evolved hearing mechanism.

millibar

A measure of pressure. One thousandth of a bar. See also dynes/cm2.

minimal distinction

See distinctive features.

MIT

Massachusetts Institute of Technology. A hive of worthwhile activity, including a chunk of speech linguistics research. Winograd did his thesis there, Chomsky is there, the Research Laboratory of Electronics (RLE) is there with the Speech Communication Group (Jonathan Allen—originator of the MITalk text-to-speech system that underlies DECTalk—and Ken Stevens still hold Professorships in the RLE), the Media Lab is there, and the Artificial Intelligence (AI) Lab is there. Situated in Cambridge, a suburb of Boston, Massachusetts.

MITalk
See DECtalk.

MIT RLE QPR
The Massachusetts Institute of Technology (MIT), Research Laboratory of Electronics, Quarterly Progress Report. Now superseded by their web-site. The speech communication group runs two sub sites. One is within the RLE tree, the other is run separately, and contains a lot more detail.

mnemonic
Strictly "mnemonic symbol". An easily remembered symbol having a clear relationship to the thing symbolised.

modiolus
The central pillar around which the spiral of the cochlea winds, and from which the bony shelf, which supports the inner edge of the basilar membrane, projects. Lying opposite the bony shelf is the spiral ligament which tethers the tectorial membrane along its outer edge. The modiolus gathers the nerves and blood vessels that serve the basilar membrane. The nerves ultimately issue forth to become the auditory nerve, which forms part of the VIIIth Cranial nerve.

monaural
Pertaining to one ear. Monotic.

monotic
A listening situation in which the stimulus is applied to only one ear. Contrasts with diotic—the same stimulus applied to two ears, and dichotic—different stimuli applied to each of the two ears. Dichotic stimuli may differ in as little as one dimension—for example, phase—or they may differ totally.

moon illusion
The illusion that the moon gets bigger as it nears the horizon. Richard Gregory, of the Psychology Department at Cambridge University, England, investigated this and other perceptual illusions. Although there is some slight refractive effect of the atmosphere that acts to increase the apparent size, the main effect is caused by incorrect (and unconscious) assumptions made by the brain in perceiving the world. High in the sky, the moon is seen as a "usual" distance, say a few score feet, and is therefore interpreted as small. As it approaches the horizon, very distant objects are increasingly seen as nearer than the moon, which therefore is taken as increasingly far away. Since it subtends the same physical angle it is interpreted as being very much larger than when high in the sky. The Ames Room effect of people changing size as they walk around an apparently normal room is equally due to incorrect assumptions. This emphasises an important aspect of perception: that perception is a highly learned skill based on what are normally unconscious assumptions about the world, and can go seriously wrong if the assumptions are violated. It is important to remember this when addressing problems of speech recognition and synthesis.

myelin
In the context of nerve transmission we have the myelin sheath, a fatty covering on nerve fibres, interrupted at the nodes of Ranvier. It accelerates the rate of propagation of nerve pulses by projecting a current field from one node of Ranvier to the next, triggering firing at the next node well ahead of the time the normal depolarisation wave would take to get there. All the faster mammalian nerves have this feature. See *nerve cell*.

nasal
To do with the nose. In this context, for Educated Southern British English (RP), it applies to nasal consonants—those in which the oral cavity of the vocal tract is closed at some point, and the velum is open, allowing energy to be radiated from the nose with a spectrum determined by the response of the vocal apparatus. The spectrum is usually weakened, especially at higher frequencies, compared to normal voiced sounds, and usually shows a split first formant peak which therefore also is reduced in amplitude. In other English dialects and in languages such as French, nasalised vowels also occur. Traditional terminal analog speech synthesisers do not do a very good job

on nasal sounds. A waveguide or tube model synthesiser with a nasal branch does a much better job because the acoustic simulation, especially the dynamics and the energy balances, are much more realistic.

NATO alphabet

Speech is made up of combinations of sounds. Such sounds tend to fall into groups which, though readily distinguished group by group, are readily confused within groups. In addition, the words used to name letters of the alphabet have many of the same sounds in common. As a result of such factors, using the common names of alphabet letters to spell words over a communication channel (where spelling is resorted to because the transmission conditions are too poor to simply speak the word) is worse than useless. The NATO alphabet is a set of letter names devised by the NATO military organization to provide a high degree of discriminability between the different names, and represents an improvement over the World War II alphabet and its predecessors. It is widely used in civilian life as an aid to clear communication over noisy channels (for example, when using ship or aircraft radio).

WW I: ACK
WW II: ABLE
NATO: ALPHA

The full NATO alphabet is:

ALPHA, BRAVO, CHARLIE, DELTA, ECHO, FOXTROT, GOLF, HOTEL, INDIA, JULIET, KILO, LIMA, MIKE, NOVEMBER, OSCAR, PAPA, QUEBEC, ROMEO, SIERRA, TANGO, UNIFORM, VICTOR, WHISKEY, XRAY, YANKEE, ZULU.

naturalness

Naturalness, applied to speech, is a complex concept. In general, synthetic-speech-by-rules is judged unnatural, and easily recognised as synthetic. The problems include at least the following: inadequate acoustics (use of a true acoustic model such as the tube model will help here); lack of detailed knowledge about how the synthesis parameters should be varied for different sounds in different contexts (and difficulty representing data and knowing which data are appropriate); inadequate models of rhythm; inadequate models of intonation; lack of knowledge on how to tie particular intonation contours to specific segment structures; and inability to choose appropriate intonation contours for specific utterances due to lack of ability to understand utterances and the intent of the "speaker". Truly natural speech-synthesis-by-rules ultimately requires that we solve the language understanding problem, and the subtleties of dialogue. In the meantime, various tricks are used, including faking understanding based on punctuation and key words, and using standard phrase structures with predefined intonation.

NCA

Noise Criteria (Acoustical). A set of octave band spectrum curves and a descriptive table are used in combination to describe an acoustic environment ranging from "Very quiet office, suitable for large conferences" (NCA 20-30) to "Communication extremely difficult, telephone use unsatisfactory (NCA > 80)". The octave band spectrum is plotted for the environment, and the number of the NCA curve which falls completely above the measured curve gives the index with which to enter the table. (See *Human Engineering Guide to Equipment Design*, ed Chapanis et al., McGraw-Hill, 1963, pp. 186-188).

nerve cell

The basic information processing unit in biological systems. Consists of a cell body (or cyton) and various kinds of processes or outgrowths. Input processes are termed dendrites and the output process is the axon. Inputs may occur directly to the cyton. Some processes (especially the axon) may reach considerable lengths. The input to a nerve cell occurs via synapses which, under certain conditions of electrochemical stimulation, may cause the cell to "fire". When a cell fires a wave of electrical depolarization sweeps from the cell into the axon and away. The depolarisation, giving rise to the so-called "action potential", is a change in the permeability of the membrane of the axon tube which allows an ion exchange between the inside and outside. Metabolic processes in the membrane subsequently restore the potential gradient ready for a new firing. The firing path may be blocked by freezing and chemicals. The processes involved are complicated and not fully understood. After firing there is an absolute refractory period when the cell cannot fire (0.4 to 2 msec in mammalian fibres), and a longer period for which the threshold of firing is raised (relative refractory period). The threshold of firing is a measure of the input activity needed to cause firing (via synapses on the dendrites and cell body). Some synapses are inhibitory, some excitatory, and it is the algebraic sum of activity (spatial summation), integrated over time with a decay factor (temporal summation), that counts. Some inputs come from specialised receptors (e.g. hair cells in the cochlea). End processes of axons are specialised electrochemical transmitters (some end on synapses to other nerve cells, some end at muscles, etc). The end process specialised to muscle operation is called a motor end-plate. A given motor end plate, on receipt of a nerve pulse, releases a chemical (acetyl choline) which first causes the corresponding muscle fibre to contract and is then almost instantly broken down again by cholinesterase -- an enzyme specialised to this function. It is now known that there are scores of these so-called "neurotransmitter" enzymes, produced throughout the body and active throughout the body. Dr. Candace Pert gives a very readable account in her book "Molecules of Emotion" (Scribner 1999) along with much other material of interest.

Some nerve gasses interfere with the breakdown process, causing the system to run amok. Nerve pulses may be observed using very fine electrodes, suitably placed. Their duration is roughly the same length as the absolute refractory period. Depolarisation of the axon may be initiated by locally applied electric currents far removed from the cell body. In such cases a nerve pulse will travel in both directions away from the site of stimulation (i.e. both dromic (normal) and anti-dromic). Speed of conduction in nerves innervating muscle fibres is conveniently measured by applying such stimulation, and then timing the duration of travel by noting the time of subsequent muscle action. Speeds vary from 2 to 100 metres per second in mammalian fibres. If an axon branches it had been thought that equal propagation down both branches was normal. Current research suggests that varying threshold, refractory periods, and the like, may allow selective propagation at such branches, depending on the time history of stimulation to the parent cell which provides yet another form of information processing in the nervous system. Cut nerves degenerate distally (away from the cell body). Regeneration usually takes place, if the cell body is undamaged, but is slow (inches per year), and does not seem to occur for a number of important nerve types (the optic nerve, for example).

nerve deafness

Deafness due to some defect in the nerves of the ear and auditory pathways. See also *conductive deafness.*

nerve fibres

The axons and/or dendrites of nerve cell.

nerve pulse

See *nerve cell.*

neuron

See nerve cell.

newton

The unit of force in the SI system of measurement, being the force required to give a mass of 1 kg an acceleration of 1 m/sec2 (gravity exerts a force of 9.81 nt on 1 kg). Abbreviated to 'N' See also *dyne*

noise

A possible unwanted component of a signal. An unwanted component that is correlated with the signal is strictly called distortion and may, in principle, be removed. For this reason it is not considered noise. Random perturbations present problems if the resulting noise frequency spectrum overlaps that of the signal since there is no easy way of removing it. Shannon showed that information can be transmitted over a noisy communication channel without error, up to a certain rate, by the using redundancy in coding.

noise cancelling microphone

A microphone so constructed that the pick-up from the talker's environment has two components in anti-phase (thus tending to cancel out), while the talker's voice is picked mainly as one component and does not cancel. Thus the signal-to-noise ratio entering the audio system is improved. The necessary phase inverter (for audio waves) is rather difficult to make of adequately wide bandwidth, so that cancellation only occurs over part of the audio spectrum. The device may also be thought of as a highly directional microphone (responding only to sounds from a particular direction). The Shure AM-10 boom microphone is an excellent example of a noise cancelling microphone, and was for many years the de facto standard for speech recognition equipment.

noise shield

A structure for providing a degree of acoustic insulation, usually between the environment and a talker's microphone (thus a pilot's oxygen mask is also designed as a noise shield and has microphone built in).

nominal scale

See *scales of measurement*.

N

See *newton*.

O/O's

observer/observers.

octave band spectrum

A description of a sound spectrum based upon the Sound Pressure Level (SPL) in a series of octave bands. Other band spectra (half-octave, third octave, etc.) are also used. If a given octave band spectrum is converted to half-octave then, since the power in each band is halved, while the reference power defined for the bands will (in general) remain the same, there will be an apparent "drop" of 3db in the overall spectrum (4.75 db for 1/3 octave). If the level is defined per unit bandwidth (Hz) rather than some form of octave basis, the apparent drop in any given unit bandwidth will be that much greater, and in general there will be an increasing drop of 3 db/octave at increasing frequencies. Of course, if, instead of merely converting, the measurements were retaken using narrower band analysis, the overall shape of the spectrum would very likely be considerably modified since there is no reason, in general, why the power in a given frequency band should be uniformly distributed throughout the band. The so-called spectrum level correction is based on the assumption of uniform distribution of acoustic energy.

oesophageal speech

Speech in which vibration of the upper part of the walls of the oesophagus is substituted for vibration of the larynx. The range and control of pitch is usually rather limited and the art requires considerable practice. When perfected, the speech can be indistinguishable (by naive listeners) from normal speech. Air is taken into the stomach and released in a controlled manner through the constricted oesophagus, providing an alternative source of excitation for the vocal tract when the normal excitation source (the vocal folds) are unable to function. The limited air capacity accounts for the characteristic pause pattern.

ordinal scale

see *scales of measurement*.

organ of Corti

A structure inside the cochlear partition, resting on the basilar membrane, and concerned with the generation of nerve pulses corresponding to displacements of the basilar membrane. The pillars of Corti, which bend over and become the reticular membrane, provide a basic framework and directly support the outer hair cells. Along the inner edge (i.e. nearest to the modiolus) are more hair cells, though the outer are four times more numerous. Other cells (Hensen's cells) also provide the body of the organ and generate steady voltage (D-C) potentials. Nerve pulses are generated by the hair cells as a result of lateral displacements of the overlying tectorial membrane acting in a direct mechanical fashion on the hairs on the hair cells. These pulses propagate down the nerve fibres that are gathered to form the auditory nerve. Apparently the outer hair cells provide a form of positive feedback gain control, and are capable of producing audible "squeals" associated with such feedback (that is, a person near the subject can hear noises coming out of the subject's ear). The entry above for *cochlea* provides to illustrations showing a cross section of the cochlear duct and a larger image of the organ of Corti.

orthography

The written form of a language. The written form cannot be spoken without a knowledge of the pronunciation, rhythm and intonation of the language. English has a particularly complex relationship between spelling and pronunciation, as may be illustrated by a poem, The Chaos, written at the turn of last century by a Dutchman (G.N. Trenite, alias Charivarius).

ossicles

These are small bones which form a jointed chain from the ear-drum to the oval window. They provide impedance matching between the pressure waves in the air, and the required displacements of the perilymph in the inner ear, by intensifying the pressure fluctuations on a 22:1 ratio. They comprise the malleus (hammer), incus (anvil), and stapes (stirrup). These are names derived from Latin, based upon their shape, which is at first sight rather curious. They are, in fact, a miracle of evolution, and not only provide near-perfect matching, but, because of their shape and attachment, protect the ear against overload. Excess movements cause rotations instead of displacements, reducing the transmission of energy to the inner ear. They also reduce bone-conducted input from the owner's vocal efforts, walking, etc. because the centre of gravity and the centre of rotation of the malleus coincide, eliminating transmission to the inner ear of vibrations of the head in a vertical direction. The oval shape of the footplate of the stapes is another evolved adaptation connected with this kind of action. A diagram provided with the entry above for cochlea shows the ossicles in relation to other auditory structures.

otoscope

An optical device designed to allow an investigator to look at the ear-drum despite the curve in the meatus (the passage between the external ear, or pinna, and the ear-drum). A source of light contained within the instrument illuminates the view.

oval window

The membrane dividing the air space of the middle ear from the fluid filled scala vestibuli that forms part of the cochlea. Attached to it by the annular ligament is the stapes, the last ossicle (or small bone) in the pressure-intensifying, noise-eliminating chain of ossicles leading from the ear-drum. Thus the oval window actually transmits the air-pressure-waves, suitably matched to the more resistant perilymph, into volume displacements of the perilymph within the scala vestibuli. A diagram provided with the entry above for *cochlea* shows the stapes which attaches to the oval window in relation to other auditory structures as well as the general location of the round window at the outer end of the scala tympani, which is partly sectioned. See also *round window*.

OVE

Pronounced "oovay" -- the resonance analog formant synthesiser constructed at the Royal Institute of Technology in Stockholm, Dept. of Speech, Music and Hearing (as it is now called) by C.G.M. (Gunnar) Fant and his associates. The original device has evolved with time, although the basic principles remain the same. OVE I was very simple having only two formants, controlled manually by a pantograph-style arm moving in a notional F1/F2 plane, with additional excitation controls. Later versions implemented a full-scale parametric, cascaded formant, resonance analog synthesiser, similar to Lawrence's Parametric Artificial Talker (PAT), with the addition of a fixed nasal resonance. Synthesisers based on series or parallel formant filters are an approximation, but only recently have the problems associated with controlling a distributed (waveguide/transmission-line/tube model), and providing the necessary data, been solved. See also *tube model*.

Pa

See *Pascal*.

palatogram

A graphic representation of the area of contact between the tongue and palate in making a speech sound. The actual areas are mapped. Moisture sensitive paper in the roof of the mouth has been used. A plate (similar to the dental support for upper false teeth) with electrodes has also been used to determine places of contact between tongue and palate on a dynamic basis.

Parametric Artificial Talker (PAT)

The first successful resonance analog synthesiser built. Invented by Walter Lawrence, of the Signals Research and Development Establishment of the British Government, in the very early 1950's.

Pascal

If taken as being a unit of SI measurement, it is a unit of pressure; 1 Pa = 1 newton/square metre. Unlike some SI units named after famous men of science, this one retains the capitalised initial 'P'.

PAT

See *Parametric Artificial Talker*.

pattern playback

A device built at the Haskins Laboratory in New York and used, during the 1950's, to determine many of the acoustic cues for speech perception, by systematic psychophysical experiments using synthetic speech stimuli generated by the device. It was based upon an acoustic domain analog of speech. It allowed representations of spectrograms (including hand painted ones) to be turned back into speech. A strip light source and cylindrical collimation system projected a line of light through a radius of a revolving wheel which turned at constant speed. Tracks running circumferentially around the wheel, and suitably spaced to correspond to harmonics 120 hz apart when transferred to a spectrogram, contained markings which modulated the light passing through each track. The rate of modulation for a given track corresponded to some harmonic of a 120 hz pitch rate, at the constant speed of rotation -- the tracks being arranged in appropriate order. The split beam of light, after passing through this modulating arrangement, contained all necessary harmonics for synthetic speech. Correct selection of those required for a given utterance was effected by reflection or transmission at the moving spectrogram. Noise was simulated by random dotting of the spectrogram in the correct frequency region. The light signals were all collected by a receiver (which therefore also summed the components), and the resulting electrical signal was converted to sound, reproducing the time waveform corresponding to the varying spectrum represented by the input spectrogram. The original paper by Cooper, Liberman and Borst, "The interconversion of audible and visible patterns as a basis for research in the perception of speech", was published in the Proceedings of the National Academy of Science, 37, pages 318-315, in 1951.

peak clipping

See clipped speech, *PPM,* and *VU meter*.

Peak Program Meter

A meter that displays the varying local peak amplitude of an input sound, rather than providing an average power measurement, like the VU meter. This is valuable as a basis for avoiding overload conditions for electronic audio equipment. A comparison is made under the VU meter heading.

percept

The mental construct built up from sensory data by a biological organism. That thing perceived by an organism. The result of perception. What is perceived may differ from what is really there because perception is based on many assumptions about the world which are quite unconscious, as far as the organism is concerned. Hence many illusions (the Ames room, the moon illusion, differing accounts of an accident by witnesses, etc.). It is interesting to note that the predominant input to the human cortex is internally generated, rather than directly from sensory input (by a factor of possibly 40:1).

perception

See *percept*.

perilymph

Watery fluid filling the scala vestibuli and scala tympani in the cochlea.

pharynx

The oral pharynx is that part of the vocal tract between the glottis and the velum. The nasal pharynx continues part way along the nasal passage.

pharyngeal

To do with the pharynx.

phase

A repeating time waveform may be considered to start each cycle at a particular time. The start time could be changed. If it were changed enough, it would simply start at what would have been the beginning of the next cycle. Phase is a description in circular measure of the relative time position of some new start time. 360 dgrees brings the signal back to its original phase. A phase angle of 180 degrees corresponds to antiphase—completely out of phase—in the case of a sin wave, for example.

phon

The unit of measurement on the subjective interval scale of loudness. The scale value is identical to the sensation level at 1000 hz, by definition, and is constructed by asking subjects to adjust a second tone to be equal in loudness to a reference tone of known sensation level at 1000 hz. A subjective ratio scale has also been constructed, for which the unit is a sone. At 1000 Hz a loudness of 1 sone is equivalent to a loudness level of 40 phons. See ***loudness***, *scales of measurement,* and *sone*.

phonation

Voicing, when producing a voiced sound.

phone

The basic unit used in describing the sounds occurring in an utterance in a language. A "Broad Transcription" requires considerable knowledge of the particular language, and produces a sequence of phonemes (also called a phonemic transcription). When trying to capture the details of a particular accent, or a new language, a so-called "Narrow Transcription" must be used, which uses phonetic symbols (defined by the International Phonetic Association or IPA) in a very general way, with a large number of modifiers (diacritics), to capture the fine detail of the utterance. This is essential for a new language, and provides the basis for determining which differences between sounds distinguish words, and which are purely distributional (that is, depend only on the particular context). See *segment*

phoneme

The modern notion of phoneme, as distinct from phone, was almost certainly first formulated by Jan Baudouin de Courtenay, a Polish philologist, in the 1870's. Daniel Jones (*The Phoneme: its Nature and Use*. Heffer: Cambridge, U.K. 3rd ed. 1967) gives an appendix on the history and meaning of the term. Phonemes only have meaning in the context of a specified language. The sounds produced in speaking the language are termed "phones". These sounds may be grouped into classes or categories. Two sounds belong in the same category if they never distinguish two words in the language. The categories are the phonemes of the language, each containing many different allophones, the variation between which is insignificant as far as meaning is concerned. Thus phonemes are functionally defined abstract categories that are specific to a particular language. The concept, though useful for describing alphabetically oriented languages, and fundamental to modern phonology through phonetics, has been a snare and delusion to those attempting machine recognition of speech based on acoustic criteria. The only practical way to view physical correlates of the phonemes is as postures of the vocal apparatus (speech postures, or articulatory targets). However these target postures are by no means always achieved, let alone maintained and are affected by context. In listening to speech, the ideal targets (or speech postures) can only be inferred from acoustic data (perhaps with some additional visual cues). They cannot, in general, be directly experienced. The fact we humans can recognise spoken language provides an existence proof for the possibility of a process that converts the acoustic waveform into a representation of what the speaker intended the listener to understand. This is the process that those trying to solve the speech understanding problem are trying to emulate. It requires the solution of some very hard problems including, ultimately, the representation and appropriate use of knowledge about the real world. See *segmentation*.

phonetically balanced word list

A word list in which the phoneme of speech occur with the same relative frequency that they do in a random sample of speech. There are certain conceptual difficulties here since these relative frequencies may vary with the kind of speech (reading aloud, conversation, etc.).

physiological analog or model

Some construct that models the functional anatomy of a biological system. Thus a physiological analog speech synthesiser represents those parts of the vocal tract anatomy that contribute to the acoustic spectrum of speech in terms of function. Such an analog may be varied dynamically by specifying physiological measurements (typically cross-sectional areas) at successive time instants. Alternatively parameters related to the positions of the articulators may be used, and the corresponding sound specification may be computed on the basis of these parameters, by converting them to the required physical realities (cross-sectional areas). Such a system would be an articulatory model. See also *articulatory synthesis*.

pillars of Corti
See organ of Corti.

pinna
Plural pinnae. The pinnae are what we refer to when we say "My ears are cold". Another term for pinna is auricle and, together with the auditory meatus, it forms a reversed megaphone to gather sounds to the ear-drum. This, coupled with the baffle effect of the head, produces a frequency dependent increase in the pressure by the time the sound waves reach the ear-drum. This is what an artificial ear duplicates. Maximum increase occurs around 2000 hz, producing a four- to five-fold pressure gain.

pitch
Often used to mean the frequency of vibration of the vocal folds (glottal rate). Strictly, it is the subjective quality correlated with fundamental frequency and measured in mels rather than Hertz. The relationship between mels and Hertz is not linear, though it is monotonic. The pitch of a 1000 hz tone 40 db above threshold is defined as 1000 mels.

pitch detection
The process of determining, by machine, the glottal rate/frequency (voicing frequency) of speech. This is not, strictly, the same as pitch, which is a subjective quality related to the voicing frequency. See *voicing detection*.

pitch synchronous analysis
See *Fourier analysis*

place theories of hearing
Place theories of hearing assert that (perceived) pitch depends on the place of maximal stimulation along the basilar membrane. By contrast the volley theory (somewhat related to the telephone theory says that pitch is perceived by a time analysis of the pattern of nerve pulses arriving from the basilar membrane. Both mechanisms are undoubtedly operative, with volley theory predominant below 1000 hz and place theory above.

PPM
Peak program meter. See also *VU meter*.

pragmatics
The study of the relation of signs to users. The study of meaning, or connotation, at a higher level than the semantic; to do with the situational and practical aspects of meaning, involving the associations of particular experiences in the world, as opposed to formal dictionary meanings. Milner's semiotic model comprises three levels: symbols (syntax); semantics; and pragmatics (C.D. Milner: *Science of Symbols in International Encyclopaedia of Applied Science*). See *syntactics*, and *semiotics*.

presbycusis
Progressive deafness, especially in the higher frequency regions (say above 3000 hz), caused by increasing age. It is possible that presbycusis is at least in part comprised of avoidable noise-induced hearing loss.

pressure spectrum level (PSL)
Is defined as the sound pressure level per Hertz, for the noise in question. If the band of frequency containing the noise is DF wide, then: PSL = SPL - 10 log DF (why not 20 log DF?) See also *intensity spectrum level.*

pretonic
See *tonic*.

probe microphone
A microphone designed to investigate sound pressure waves in restricted areas. The active area of the microphone (that part which is affected by incident sound waves) is made as small as possible and the microphone is constructed to allow the active area to be placed in difficult-to-get-at places.

process (of nerve)
A thread, filament, fibre or what have you that extends from a nerve cell body.

proclitic

"Proclitic" is a term applied in the context of grammatical structure. In relating syllables to rhythmic structure in spoken language, some syllables are grammatically related to those preceding and some to those that follow. The former are called "enclitic" syllables, the latter "proclitic"syllables. Jassem regards proclitic syllables as analagous to anacruses in music, and does not include them in the basic units of his analysis of rhythmic units (which he calls "rhythm units").

prominence

In speech this refers to the degree to which a syllable or sound subjectively stands out from its neighbours. It is related to four factors, according to Daniel Jones: basic sound (or syllable nucleus) quality, quantity (duration), stress (acoustic intensity), and intonation (pattern of change of glottal vibration frequency). Instrumental and psychophysical approaches confirm the importance of these factors and rank stress, quantity and intonation in that order as having increasing effect on prominence. In fact, intonation and duration are probably the most powerful subjective correlates of prominence—that is prominence is a subjective effect on the listener as opposed to stress (one correlate of prominence) which is a subjective activity on the part of the speaker. See also *salient* and *syllable*.

proprioception

The feeding-back of information concerning the state of joints, muscles etc. in a biological system to the parts of the nervous system concerned with body image. It is proprioception that underlies much of our kinaesthetic sense, or sense of where the various parts of our body are in relation to one another, and absolutely. There are special receptors (muscle spindle organs, receptors on tendons, etc.) which transmit specific proprioceptive feedback.

prosody

Slightly tricky to define. Basically a suprasegmental feature (at a higher level than the segmental) but, broadly (following Lehiste—*Supra-segmentals*: MIT Press 1970)—those aspects of speech that are other than segmental in character, in the sense that they overlay the segmental aspects and form patterns extending in time, or intensify phonetic factors already present to a lesser degree. Thus, taking pitch, stress and (time) quantity as the prosodic features of English, it is seen that while voicing is an inherent feature—identifiable at a moment in time, glottal rate is an overlaid function. The fundamental frequency serves both to identify a segment as voiced and to contribute toward the intonation pattern of the utterance. Each segment requires some duration, but rhythm is the formation of time patterns by the manipulation and concatenation of inherent duration, "beats" occurring on strongly stressed (and therefore salient or prominent) syllables. Each segment has some inherent subjective "loudness" but, though this is mediated by a combination of intensity, duration and pitch variation, prominence is a result of reinforcing or diminishing qualities already present, on a contrastive basis, in time.

psychological "set"

The expectation a person has which tends to influence his or her perception of a situation or stimulus.

puff screen

A device, often a plastic foam cap, attached to a microphone to prevent the unpleasant acoustic effects (and possible damage) caused by puffs of air from the talker's performance striking the sensitive parts of the microphone. A foam cap also protects the microphone against moisture from the breath and against drops of saliva.

quefrency

See *cepstral analysis*.

ratio scale

See *scales of measurement*.

receptors

Organs (devices) for accepting input from the environment. Sensory data receivers.

redundancy

The property a structure (physical, informational, ...) has when there are components which duplicate the role of other components in any way. If a component is completely redundant, it may be removed without loss of integ-

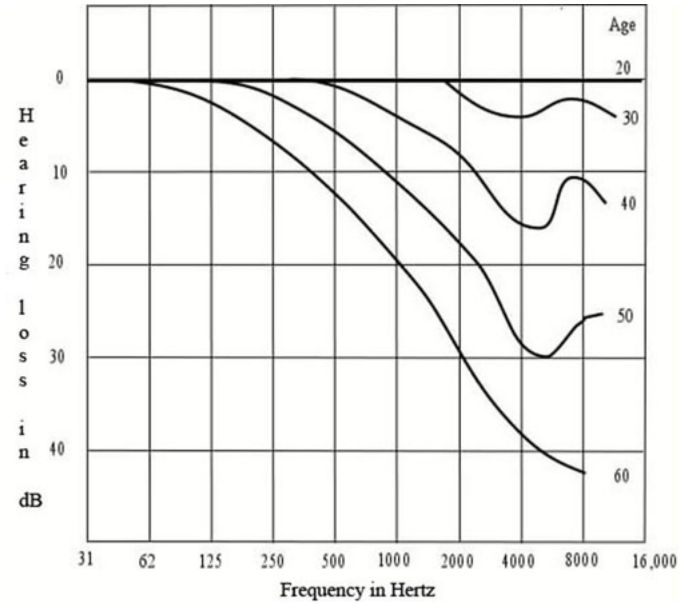

Figure 13: *Hearing loss plotted by age*

(Hearing loss is plotted for five different ages; each shows the average decibel loss by frequency compared to a 20-year-old with normal hearing.)

rity to the structure. In physical structures, this provides protection from collapse due to damage. In an information structure, components of the "message" may be lost while still retaining all the information. In information theoretic terms redundancy is expressed in terms of the conditional probabilities of joint events. A message with redundancy has more physical symbols than are necessary to represent the formal information content and will still convey the whole original information correctly even if somewhat damaged by noise. Claude Shannon (Bell System Technical J. 1948: *A mathematical theory of communication*) showed that it is possible to transmit information without error, even in the presence of noise, provided the rate of information transmission does not exceed the well-defined capacity of the channel. The trick is possible because of the use of redundancy, and the work spawned a major research field concerned with coding theory for purposes of error-free transmission of information. See *information*.

refractory period
See *nerve cell*.

Reissner's membrane
An insulating membrane forming one wall of the triangular duct that partitions the spiral space inside the cochlea into two parts, (the scala vestibuli and scala tympani), which are joined only at the helicotrema. The other walls are the basilar membrane and the outer shell of the cochlea.

Reiz
See *difference limen*.

resonance analog
An acoustic domain model of speech production based upon the known resonant behaviour of the vocal apparatus and the known characteristic of the exciting source. In producing synthetic speech, parameters describing the varying resonances and excitation function are supplied to a set of filters (in cascade, or in parallel), which reproduces the speech wave on the basis of the controlling parameters, and the constraints inherent in the embodiment of the filter model. Lawrence's PAT (Parametric Artificial Talker) is a typical cascade resonance analog synthesiser. Holmes used a parallel filter model. The advantage of the former was that fewer controls were required, because the amplitude adjustment occurred automatically. The advantage of the latter was that more accurate spectral modelling was possible, at the expense of additional formant amplitude controls. Lawrence's machine was the first successful resonance analogue speech synthesiser when it toured the USA in 1953. See also *acoustic analog, physiological analog* or model, *tube model* and *terminal analog speech synthesiser*.

resonance theory of hearing
A place theory of frequency analysis in auditory perception that assumes different sections of the basilar membrane are tuned to different frequencies and respond like a mechanical frequency analyser, transmitting the place

of maximum vibration to the central system and thus marking the frequencies present in the stimulus. The known properties of the basilar membrane render the theory untenable, but in practical terms it differs little from the accepted travelling wave theory which is another "place" theory, but one that is generally accepted as part of the mechanism by which frequency content is discriminated. See also *volley theory*.

resonance vocoder

See *vocoder*.

resonator

A system or structure that stores oscillatory energy at particular frequencies (often only one) that depend on its physical properties and the way energy is supplied to the system or structure.

reticular membrane

Part of the organ of Corti which supports the outer hair cells.

rhamonics

See *cepstral analysis*.

rhythm units

The name given by Jassem to his formulation of rhythmic units. See *rhythmic units, foot* and *isochrony/isochronicity*

rhythmic units

In a stress-timed language such as English, the speech may be divided into rhythmic units related to the occurrence of the stresses, or beats. M.A.K. Halliday, following David Abercrombie (one-time Professor of Phonetics and Linguistics at the University of Edinburgh), puts the rhythmic unit boundaries just before each syllable bearing primary stress and calls the unit a "foot". Jassem's scheme (he was at the Polish Academy of Science, Poznan) is similar, but excludes proclitic syllables from consideration, likening them to the anacruses in music. Such frameworks can provide a basis for workable models of English rhythm, but require extensive analysis of real speech segments. One such study by the author of this conceptionary, together with Jassem and Witten, slightly favoured the Jassem formulation, but a subsequent model of rhythm for use in speech-synthesis-by-rules used the Halliday formulation because it is a little simpler.

RIT

Also KTH (from the Swedish name). The Royal Institute of Technology in Stockholm, where C.G.M. Fant had his Speech Technology Laboratory (now the Dept. of Speech, Music and Hearing). One of the two or three most important speech research laboratories in the world, and certainly as much a pioneering institution as the Haskins Laboratory. Like most institutions these days, their research is accessible through their websites. Their regular paper reports have been discontinued (though papers still appear in the peer reviewed journals).

RIT STL QPSR

Royal Institute of Technology (Stockholm) Speech Transmission Laboratory Quarterly Progress and Status Report. See RIT.

RL

See *difference limen*.

RLE

The Research Laboratory of Electronics, at Massachusetts Institute of Technology. Much important speech and language research has been carried out there. Dennis Klatt was there; Jonathan Allen (originator of DECtalk/MIT-alk) and Ken Stevens still hold professorships there in the speech communication group.

rms

"rms" stands for "root mean square". Typically encountered in measuring things like alternating electrical voltage. The rms value reflects a usable average electrical pressure (voltage). A simple minded approach would say the average voltage was zero. By using an average measure that involves squaring and taking the square root, the apparent cancellation of positive and negative halves of the waveform is avoided, and a measure of power can be computed.

round window

One of two membranes dividing the air space of the middle ear from the fluid filled scala tympani. Without this flexible end to the scala tympani there would be no displacement of the inner ear fluids and no displacement of the basilar membrane. It is complementary to the oval window. See *oval window.*

S/N ratio

See signal to noise ratio.

S/S’s

Subject/subjects.

sagittal section

Section taken through the plane dividing an organism into right and left halves.

salient

In a stress-timed language, such as English, where primary stressed syllables seem to fall at fairly regular intervals, regardless of the number of unstressed syllables intervening, the rhythm may be regarded within a framework of feet, like the bars in music. Each bar either begins with a stressed beat or a silent beat and, in Abercrombie’s terminology (which M.A.K. Halliday follows), the former is called a salient syllable and the latter a silent stress or phonological pause. There is little if any acoustic evidence for the second. See *stress, prominence* and *isochrony*.

scala vestibuli/scala tympani

The two longitudinal chambers of the tube which, coiled up, forms the cochlea, separated by the cochlear partition—itself a fluid filled duct. The scala vestibuli starts at the oval window (to which is attached the stapes ossicle by the annular ligament) and ends at the helicotrema where it joins to the scala tympani. The scala tympani extends between the helicotrema and the round window. The cochlear duct ends at the helicotrema and does not communicate with either the scala vestibuli or the scala tympani. See *cochlea* and associated diagrams*, bony labyrinth* and *membranous labyrinth.*

scala media

The interior space of the cochlear partition—the cochlear duct, filled with endolymph. There is no fluid connection with the perilymph-filled scala vestibuli or scala tympani. See *scala vestibuli/scala tympani* and *cochlea.*

scales of measurement

Stevens (*Handbook of Experimental Psychology*, Wiley, 8th printing 1966, pp 22 et seq.) gives a good discussion of scales of measurement. Measurement is the assignment of numerals to objects or events according to rules. There are four generally accepted scales of measurement. Nominal (which simply labels classes and only permits counting of events and determination of mode): Ordinal (which rank orders classes, is preserved by any monotonic increasing functional transformation, preserves the notion of “between”, but whose interclass intervals are of unknown relative size); Interval (a commonly encountered scale for which most statistics are valid, for which intervals of equal magnitude really represent equal distances in terms of the quality being measured, but which have an arbitrary zero—measurements on one interval scale may be transformed into intervals on another by an (x’ = ax” + c) type of transform, as for Celsius and Fahrenheit temperature scales); and Ratio (for which it makes sense to say that a quantity which is numerically twice as big as another really does represent double the quantity—conversion from one ratio scale to another is simple multiplication by a constant, (x’ = ax”) and there is a real hard zero.). Psychology aspires to interval scales but often achieves only ordinal scales to which it applies invalid statistics with surprising success. Physics depends on ratio scales.

SEF

See *single equivalent formant*.

segment

The result of applying segmentation to speech. Traditionally a speech segment corresponds to a phoneme and may be associated with a speech posture. However, a segment is more accurately identified with a phone. Phones falling in a given phoneme category vary widely. This fact, and the difficulty of providing a reliable algorithm for deciding where boundaries between segments are located cause great difficulty for simple approaches to speech recognition by machine.

segment synchronisation

An approach to solving the segmentation problem for speech recognition purposes. Early work in America done by Vicens under the supervision of Reddy gives the method its name. The basic idea is to match input sounds against stored templates of sounds in the machine's vocabulary on the basis of the statistics of the sound segments at the word level. However, an initial analysis classifies words on the basis of only a few easily detected kinds of segment (silence, friction, vowel) and portions remain unclassified, but statistically described. When an unknown sound is input, only those stored templates having the same broad pattern are considered for matching. The known segments are mapped between the unknown sound and a possible matching template, which also puts the unclassified segments into the most plausible relationship. All segments are then compared on a detailed statistical basis (some segments, say those unrepresented in one sound but existing in the other, will contribute negatively). The best fit from all the candidates is then chosen as the identification response by the machine, thus "recognising" the input word. Modern methods using Hidden Markov Models (HMMs) have formalised and replaced this approach. Though simpler to implement, in some sense, it is not clear that HMMs achieve all the advantages of Vicen's approach. Both approaches take what I call a "salami" approach to speech (they slice the speech into pieces at what are inevitably arbitrary time boundaries, and ignore the asynchronous nature of the acoustic cues to speech perception). See also *segmentation*.

segmental level

The level of analysis, description, etc. that attempts to deal with sound segments (phones). See *phone, segment,* and *segmentation*

segmentation

The division of continuous speech into successive chunks (the "salami" approach), each associated with a single phoneme. Ken Stevens of the RLE at MIT wittily and correctly described the problem of doing this as "the problem that you can't". Nevertheless it is still attempted on the basis of assumptions that have limited validity in phonetics, and less validity in speech recognition. The problem is inherent in the definition of what is a phoneme, as well as the fact that actual speech sounds (phones) run into one another and influence one another. See also *time normalisation*.

semantics

The study of the relation of signs to objects (S.S. Stevens). The study of meaning and the patterns of meaning in a language. Semantic constraints in a language exist at a higher level than syntactic constraints. Thus "The boy ate mount Everest for now" is syntactically impeccable, but semantically wrong, while "Ate mount for boy the now Everest" is wrong on both counts. Constraints imply redundancy which implies the possibility of correcting for errors or inadequacy at lower levels of analysis, which is why semantic and syntactic constraints are important in speech recognition and communication in noise. See *pragmatics, syntactics* and *information*.

semicircular canals

Parts of the bony labyrinth concerned with balance. They interconnect with the scala which form part of the space inside the cochlea. The cochlea is concerned with hearing. See *bony labyrinth* and *membranous labyrinth*.

semiotics

The study of signs and symbols in various fields. The study of signalling systems. Comprises syntactics, semantics and pragmatics. See *information*.

sensation level

Not to be confused with threshold of sensation. The sensation level of a tone is the number in decibels by which it exceeds the threshold of hearing for that frequency. SL = 10 log (I/It) db re It—the intensity threshold.

sensors

Organs (devices) for receiving input from the environment. Sensory data receivers (where "sensory" comes from the same root as senses—those faculties of perception possessed by biological organisms). Receptors.

SI

Systeme International (SI) is a standard metric-based measuring system adopted in 1960 at the 11th General Conference of Weights and Measures. It is based on metres, kilograms, seconds, amperes (units of electric current), degrees Kelvin (which measure temperature -- degrees Kelvin = degrees Celsius + 273), candelas (which measure luminance), and moles (which measure the amounts of substances, and relate to the number of molecules).

sign

Used in linguistics to indicate the way that linguistic expressions (words, phrases, ...) represent the situations, objects, and so on for which they stand. Swiss linguist de Saussure strongly affected relevant discussion in linguistics by contrasting the signifier with the concept signified. Semiotics is the science of signs. Terms such as sign language use "sign" in a very restricted sense.

signal to noise ratio

A measure of how the signal level compares to the noise level. It is expressed as the power of the signal relative the power of the noise, both in db, and is thus dimensionless.

SIL

See *speech interference level.*

single equivalent formant

By processing the speech spectrum, Charles Teacher and his associates at Philco-Ford produced this measure of speech spectra for speech recognition purposes. In essence it follows formant 1 in back vowels and formant 2 in front vowels, providing a single dimension for the discrimination of vowel sounds. See formant

Sonagram

See *sound spectrograph.*

sone

The unit of measurement on the subjective ratio scale of loudness. One sone is defined as the loudness of a tone which is 40 db above the threshold of hearing at 1000 hz, i.e. the loudness of a 1000 hz tone at a sensation level of 40 db is 1 sone. The scale was set up by asking subjects to adjust the ratio of two tones presented simultaneously. It has a real zero which follows the threshold of hearing as the frequency varies. A loudness of 2 sones sounds twice as loud as one of 1 sone. Another subjective unit of loudness, on which an interval scale is based, is the phon. See *loudness, scales of measurement,* and *phon* with the accompanying diagram.

sound intensity level

Abbreviated as IL. Defined as: IL = 10 log (I/Io) db re. Io watts/m2

sound pressure level

Abbreviated as SPL. Defined as: SPL = 20 log (p/po) db re. po nt/m2

sound spectrograph

A device for carrying out a frequency analysis of audio signals and presenting the results as a pattern of time-energy-frequency variation called a spectrogram. The first such apparatus was constructed by Dunn and others at Bell Laboratories in the early '40's and was subsequently produced commercially by the Kay Electric Co. of Pinebrook, New Jersey, under the brand name of "The Kay Sonagraf". The machine used Teledeltos paper (which darkens when an electric current is passed through it) to record the output which Kay called a Sonagram. The machine used a mechanically tuned filter linked to a motor-driven scanning/marking head, which was moved, axially,up the outside of a cylinder around which the recording paper was wound. The cylinder revolved, generating the time scale as an x-axis. Darkness of marking indicated the energy present, while frequency was displayed up the paper as the y-axis, due to the mechanical linkage between frequency of analysis and the scanning system. The sound signal being analysed was recorded around the periphery of a disc which was locked on the same axle as the recording drum. It was able to take spectral sections, generate amplitude displays, and vary the frequency scale and analysing filter bandwidth to suit formant or pitch analysis. An extra attachment allowed "contour" spectrograms to be produced which added intensity contours to the fine grey scale energy display. Purely electronic equipment has replaced the original electromechanical Sonagraf, and these are in the process of being displaced by suitable software running on general purpose desktop and laptop computers. The new media have yet to duplicate the subtlety of marking that made the Sonagram such an excellent source of speech data. However, they are more flexible, and avoid most of the tedious calibration and maintenance problems of the early equipment. The figure shows a spectrogram of the vowels that are characteristic of General Amercian (GA).

source-filter model of speech production

The speech production process is modelled as a "source", or excitation function (periodic, random, or a combination of both), which generates the spectral fine structure): and a "filter" or vocal tract response function, characterised by the values of resonances (formant peaks), anti-resonances, losses, and radiation impedances. The source function may be defined crudely by 4 parameters (Ax, Fx, AH1 and AH2) and the filter by another 4 (formants F1 to F3 and a fricative noise spectrum shaper). When these act together, a varying spectrum with the correct shape and fine structure may be generated by a synthesiser based on such an approximation to the source-filter properties of the vocal system. A resonance analog speech synthesiser is thus based on the source-filter model of speech production. The Klatt software synthesiser is such a synthesiser, but implements a far more detailed realisation of the model. The Parametric Artificial Talker (Lawrence 1953) was the first successful, true source-filter-based synthesiser. See *transmission line analog speech synthesiser* and *tube model*.

spectral envelope

The broad shape of the (sound) spectrum, as might be revealed by a wide band analysis (300 hz bandwidth) using a sound spectrograph. The chief determiner of the spectral envelope of speech sounds is the filter function of the vocal tract. The spectral envelope often exhibits peaks and these are termed "formant". See also *voicing*.

spectral fine structure

See *voicing*.

spectrogram

See *sound spectrograph*.

spectrographic analysis

See *sound spectrograph*.

spectrum

See *frequency spectrum*.

spectrum level

See *intensity spectrum level*.

spectrum level correction

A sound spectrum may be described by dividing the frequency spectrum into an arbitrary set of frequency bands and expressing the energy present in each band as the (sound) Intensity Level (IL) or sound pressure level (SPL). In general, assuming uniform impedance, and appropriately chosen reference levels, these are the same. If the chosen bands are 1 hz wide, the resulting description is termed the Intensity Spectrum Level (ISL). If bands of any other width (say ΔF) are used, the spectrum may be converted to an ISL description by applying a correction factor: -10 log ΔF to each IL or SPL value taken in the various bands. If the original description was in terms of (say) an octave band description, then the correction factor puts each successive band down by 3 db more than the last, since each band is twice the width of the last, as frequency increases, and, for octave bands of equal power, the intensity per Hz is halved (-10 log 2 is -3). See *intensity spectrum level, octave band level* and also visit:

https://www.usna.edu/Users/physics/ejtuchol/documents/SP411/Chapter8.pdf

speech interference level

An approximate method of predicting the intelligibility of speech is based upon the speech interference level, and is valid where the noise spectrum is reasonably continuous (ventilation noise, engines, ...). The SIL is the arithmetic average of noise in the three octave bands 600-1200, 1200-2400, and 2400-4800. Given the SIL a table is entered which gives the maximum distance at which 75% of phonetically balanced words (98% of test sentences) will be heard correctly. This degree of intelligibility corresponds roughly to an articulation index of 0.5.

speech posture

A target for the vocal apparatus in human speech. In approximating a succession of speech postures, with appropriate concurrent control of breath and pitch, the human vocal apparatus produces the sounds of speech. The author prefers to talk about speech postures as a basis for machine synthesis of speech by rules, because phones are the sounds of speech—a result rather than a cause, and phonemes are abstract categories of sounds which may be produced by a variety of speech postures.

speech synthesiser

A speech synthesiser models some aspect of the speech signal—its form, production, etc.—in order that speech sounds (phones) may be produced artificially from suitable control signals. The control signals required for an utterance may be extracted from analyses of real (natural) speech, leading to "compressed speech" based synthesis. The signals may also be computed on the basis of general knowledge about speech cues, articulatory constraints, etc., without reference to a particular natural utterance, and intelligible speech still produced. This latter method of synthesis is termed speech-synthesis-by-rules. Although the term has been applied to segmental level synthesis, speech completely by rule must include algorithms for producing the acoustic correlates creating the suprasegmental aspects (rhythm, intonation, stress, and so on), and even converting ordinarily spelled speech into strings of sound symbols first (text-to-phonetic-symbols). All of these levels (including segmental synthesis) are still very much research topics, but most progress has been made at the segmental level. Converting normal orthography to spoken words, by machine, is the "text-to-speech problem". The first text-to-speech system was created by Holmes, Mattingley, and Shearme (Speech synthesis by rule, *Language & Speech* 7 (3) July-September 1965)

speech-synthesis-by-rules

The production of speech by machine, without direct reference to any original utterance, based on general knowledge of speech characteristics, together with models of the acoustic behaviour of the vocal tract and related speech production essentials, rhythm, and intonation. See speech synthesiser.

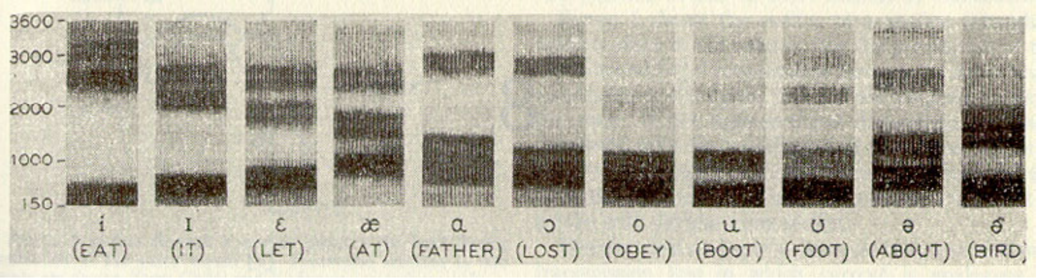

Figure 14: *Broad band spectrograms of the eleven vowels of General American (intensity against frequency and time)*

(From: The calculation of vowel resonances, and an electrical vocal tract by H.K. Dunn 1950, J. Acoust. Soc. Amer., 22, pp 740-753)

speed of sound

The speed of sound, c, in air is given by:

$c = \sqrt{((\gamma.p)/ \rho)}$ metres/second

where γ (the ratio of specific heats at constant pressure to that at constant volume) is 1.4 at 20°C, p (static pressure) is in newtons/square metre , and ρ (density) is in kg/cubic metre. At 20° Celsius and standard pressure, the speed of sound is about 343 m/sec in air, and varies as the square root of the absolute temperature (about .63 m/sec/°C at room temp).

See the Wikipedia article on the topic for more details.

SPL

See *sound pressure level.*

stapes

A very small stirrup-shaped bone forming part of the chain from the ear-drum to the oval window. See *ossicles*.

stop sound

See *voiced stop*.

stress

Daniel Jones describes stress (in speech) as the degree of force with which a sound or syllable is uttered, and says that it correlates with the acoustic intensity. He distinguishes stress from prominence which, he says, depends on combinations of quality (some vowels are inherently more prominent than others), quantity (duration), stress and (for voiced sounds) intonation (the pattern of glottal frequency variation). He defines a syllable as a sound sequence having a small peak of prominence and notes that this may be increased or decreased by changing length, stress, or intonation, which are closely inter-related in English. Stress and intonation modifications are usually combined to cause prominence. Instrumental studies of spoken English confirm the importance of all these factors and suggest that, while many writers discuss levels of stress in spoken English (word stress which may be unstressed, primary stress or secondary stress; and sentence stress), they would be more nearly correct to talk about levels of prominence, which is the important perceptual effect that we are really concerned with. This would avoid the considerable confusion that has been generated in the field, and keep faith with Jones. See *tonic*.

subjective

As perceived by a human rather than as measured by instruments. Instrumental measurement, in theory, leads to "objective" measurements. It is worth noting that a person's perception of instrument readings is itself subjective; a well designed instrument gives repeatable readings for the same physical phenomenon regardless of the person who operates it—though there remains the problem of interpretation.

Well-trained people can supply highly repeatable subjective results (e.g. for testing communication systems). The real difference between objective and subjective measurement may be illustrated as follows. If a well designed instrumentation system and a well trained team of humans are set the task of quantifying the same physical phenomenon, using a standard reference point (e.g. the loudness (subjective measure) and intensity (objective measure) of a standard tone are agreed to be the same), they will NOT agree on the ratio between other values of the stimulus and the reference, even though they will very likely agree on the the rank order (ordinality) of the other stimuli on their respective scales. In other words, subjective and objective scales measuring the same phenomenon are not linearly related. Some objective scales relate better to subjective than others (e.g. decibels to loudness). See *scales of measurement*

subjective testing

The basic method of measuring speech intelligibility in communication systems. Requires highly trained test subjects, and careful experimental procedures. Any method of evaluating speech intelligibility in a system which does not use proper subjective testing is either an estimate (based upon instrumental readings and assumed relationships—see, for example, articulation index), or is to be regarded with grave suspicion.

superior olive

See *auditory pathways*.

supraglottal

Above (higher than, towards the mouth from) the glottis, which is the opening between the vocal folds which are situated within the larynx.

suprasegmental

Literally "above the segmental". Features extending over more than one speech segment. See also *prosody*.

syllable

"Syllables" are not well defined, though even children seem to have little difficulty detecting syllables in spoken English, so that the meaning of the word is often taken for granted. A syllable must have a nucleus (usually a vowel or diphthong, though the function may be taken on by some consonants such as /n, l, z/). The inherent intensity of such sounds, coupled with variations in pitch, often with weaker consonants helping in separation between them, lead to the perception of the "lumps" in speech that we call syllables. The consonants occurring between nuclei have a definite sense of belonging to one nucleus, rather than the other, though this belonging can be reversed by marked changes in the relevant segmental durations, including the the introduction of pauses, or the occurrence of stress. See also *juncture* and *prominence*.

syllable nucleus
See *syllable*.

symbol
Something which stands for something else. It is usually more accessible than what it stands for, and usually has some obvious relation to the thing symbolised. Symbols thus have meaning (denotations or semantics), unlike signs which are related by syntax. A "mnemonic" symbol is easily remembered.

synapse
An electro-chemical junction between the axon of one nerve cell and a dendrite or the cell-body (cyton) of the next. Unlike a nerve fibre, a synapse will only transmit in one direction. Synapses may be inhibitory as well as excitatory.

syntactics
The study of syntax, the relation of signs to signs. The study of the rules for compounding words into sentences and utterances of languages. The study of grammar. See *semantics, pragmatics* and *semiotics*.

talkers side tone
The fraction of speech input that is fed back to the ear(s) of a talker in a communication system. It ensures the system appears "alive" to a user and may be manipulated to influence how loudly a talker speaks. Delayed feedback can slow a talker down, and may disrupt the talker's speech altogether if the delay is around 200 milliseconds.

tectorial membrane
The stiff membrane that lies over the hair cells of the organ of Corti and is displaced laterally with respect to these as the basilar membrane arches up and down under the influence of transmitted pressures in the lymph ducts above and below the cochlear partition. This lateral displacement activates the hair cells and transmits nerve pulses along the auditory pathways and ultimately to the auditory cortex. The combination of the relatively stiff tectorial membrane and flexible basilar membrane, which act together to stimulate the hair cells by shearing forces, as the basilar membrane is displaced by pressure gradients across it, provides a near perfect match between the fluid resistance in the perilymph and the much more resistive mechanical system of the hair cells. Taken together with air-to-fluid matching provided by the ossicular chain, this explains the incredible sensitivity of the human ear, which approaches the theoretical limit. Displacements of the ear-drum, at the threshold of hearing, correspond to about the diameter of a hydrogen atom! Some subjects may actually hear random motion of air molecules, under optimum conditions, as part of the base level noise in their system. See *threshold of hearing, watt* and *ossicles*.

teeth ridge
See *alveolar ridge*.

telephone theory of hearing
Assumes the ear has the role of simple transducer, rather than one of frequency analysis, only converting the sound pressure wave into pulse trains in the nervous system whose time characteristics preserve the time characteristics of the original pressure fluctuations. The frequency analysis is assumed to be performed by some (unknown) central nervous system mechanism. Central analysis of pulse frequency undoubtedly plays a role in auditory perception, but peripheral tone discrimination also plays an important role. See *place theories of hearing* and *volley theory*.

terminal analog speech synthesiser
A lumped property black-box style of electronic circuit that is designed to reproduce the transfer-function of the vocal tract without regard to the distributed properties (physical properties) the tract is known to possess. A resonance analog speech synthesiser is one such terminal analog speech synthesiser. A physiological analog on the other hand is called (by Flanagan, of Bell Labs who originated the name "terminal analog") a transmission-line analog. A physiological parameter speech synthesiser might be either a terminal analog or a transmission line analog depending on how the physiological parameters were turned into speech, according to Flanagan. In all cases (acoustic analog, acoustic parameter or resonance analog, physiological analog, physiological parameter analog, terminal analog, and transmission line analog) the derivation of sound from the input parameters could be done

by: special-purpose analog hardware; special purpose digital hardware (e.g. a Digital Signal Processor or DSP); a general purpose analog computer; or a general purpose digital computer. Whichever means was used, some underlying model of the speech production process would be essential. In practical terms, formant synthesisers are terminal analogues and articulatory synthesisers are distributed analogues (transmission-line analogs, waveguide, or tube models) because the conversions that would otherwise be required are too hard, in the present state of knowledge and technology.

terminal limen
See *difference limen*.

tetragram
Four letters appearing together, in order. A study of tetragram frequencies may be of value in automated language processing. A list of all legal tetragrams appearing in Webster's dictionary was published in the mid-sixties through the US Armed Services Technical Information Agency (ASTIA).

theories of hearing
See *volley theory, place theory, resonance theory, travelling wave theory* and *telephone theory*.

third octave band spectrum
See *octave band spectrum*.

threshold of hearing
The response of the human ear coupled to the brain is a purely subjective quantity and cannot be measured directly, nor directly related to objective physical measures, by any simple or universal function. Hence there are objective measures (frequency, intensity, ...) and subjective (pitch, loudness, ...) which have different units. A common method of expressing the objective threshold of hearing is as the acoustic intensity level with respect to the audiometric zero (2 x 10^{-5} nt/m^2). At this level, on average, pure tones may just be detected by subjects. The acoustic intensity level is plotted against frequency and shows the variation in threshold as a function of frequency. Although the range extends from about 20 to 20000 hz, the ear is most sensitive from about 1000 to 4000 hz, where some subjects can hear tones below the standard audiometric zero. The normal threshold range for young adults is around 30 db and the acoustic intensity at which the most sensitive human ear can detect sounds approaches 10-14 watts/m^2. Audiometric zero represents an acoustic intensity of 10-12 watts/m^2. See *threshold of pain, presbycusis, watt*, and *audiogram*.

threshold of pain
As the intensity of sound increases, there comes a point where subjects report "discomfort", "tickling", "pricking" etc. The level varies somewhat according to the frequency and experimental design, but not as greatly as does the threshold of hearing. Higher mean levels may be tolerated for clipped speech than are tolerated for unclipped speech. The maximum tolerable level is variously described as the threshold of sensation, threshold of pain, etc.

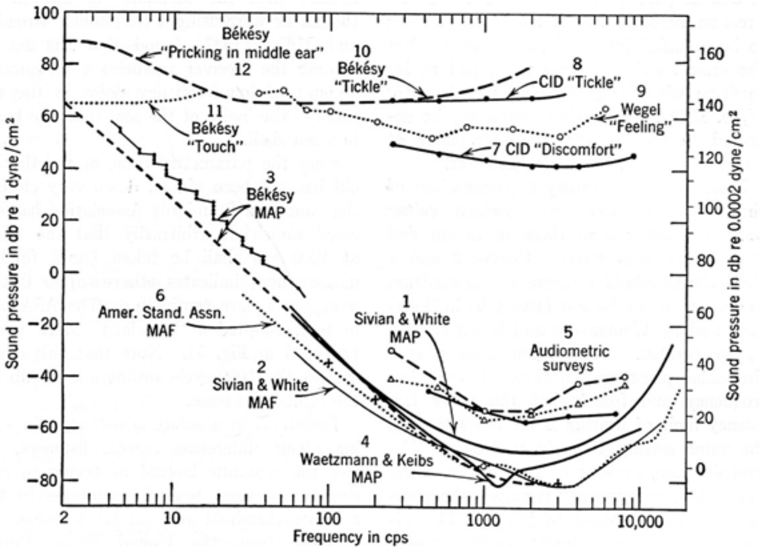

Figure 15: *Thresholds of hearing, including "pain" (A repeat of Figure 10)*

Page 995 (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

Hearing damage may still result from levels which are lower than this. The following figure summarises a number of studies of various thresholds at the lower and upper limits of hearing. See also *hearing loss*.

threshold shift
The difference in db between the threshold of hearing for a tone and the level of the tone at which it is just audible in the presence of a masking noise. This difference is used as a measure of the masking effect of the noise. If the difference is small, the noise does not mask the tone very effectively. If the difference is large, the noise is an effective masking noise for the tone.

time domain
The class of representations of a phenomenon comprising functions of time. A speech waveform is in the time domain. This contrasts with the results of a Fourier analysis of the waveform which would give a frequency domain representation (a frequency spectrum). A time waveform comprises a distribution of amplitude against time. A frequency spectrum comprises a distribution of amplitude against frequency. They bear an inverse relationship. The procedure for going from the first to the second being the inverse of going from second to first.

time normalisation
A process which, if successful, would convert speech segment durations to some standard value so that comparisons between input sounds and reference sounds could easily be carried out for purposes of speech recognition. Dynamic Time Warping (DTW), first proposed by Velichko and Zagoryuko in the early '70's, is a technique for partial time normalisation in which reference segments may map onto more than one unknown segment, or vice versa, in order, with some segments perhaps being omitted. Hidden Markov Model (HMM) algorithms inherently embody a similar kind of time normalisation insofar as the states of the model may be visited repeatedly or omitted. If time normalisation could be perfected, segmentation would be unnecessary. Conversely, if segmentation could be perfected, time normalisation would be unnecessary. They are different aspects of the same problem.

TL
Terminal Limen. See difference limen.

tonality
The division of an utterance into "tone groups", for purposes of intonation analysis or assignment, according to M.A.K. Halliday's scheme of things. See *tonic*.

tone
Has three meanings. The common meaning refers to a sound comprising a single sinusoidal signal. In M.A.K. Halliday's model of British English intonation, it refers to the assignment of a particular intonation pattern to a particular tone group (composed of one or more feet). In other systems the analogous attribute is variously called "tune" (Kingdon & others), "tone- patterns" (Palmer & Blandford), "pitch morphemes" (Wells, though he says little about the meanings!), "intonation contours" (Pike, who treats them somewhat as supermorphemes) and so on. See tonic. Tone is also a lexical feature of tone languages such as Mandarin, where words can be distingushed on the basis they have different "tones" independent of suprasegmental pitch variation.

tonic
In M.A.K. Halliday's formulation of the intonation of British English (*A course in spoken English: Intonation*, Oxford U. Press 1970) utterances are broken into feet—analogous to bars of music—beginning with a stressed syllable. These feet are then grouped into tone groups (tonality), assigned tonic stress (tonicity) and assigned an intonation pattern type (tone). The tonicity divides the tone group into the pre-tonic and the tonic, with the first syllable of the first foot of the tonic being the tonic syllable, carrying tonic stress—making it the most prominent syllable in the tone group. The tonic marks the information point of the tone group which corresponds, usually, to a clause or sentence.

tonicity
The division of a tone group into pre-tonic and tonic. See *tonic*

transducer
A device, often mechanical, for transforming a physical measure into another form, for some purpose. Thus a pressure transducer might produce an electrical voltage proportional to the applied pressure.

transmission line analog speech synthesiser

A model of the vocal apparatus used for speech synthesis that models the distributed properties of the vocal apparatus (that is, its tube-like properties). Flanagan contrasts it with a terminal analog. A physiological analog is an example of a transmission line analog. See also *terminal analog, articulatory synthesis*, and *tube model*.

travelling wave theory of hearing

The currently accepted theory of frequency analysis by the mechanical parts of the human auditory system. Because of the mechanical set-up in the inner ear, and the hydrodynamics of the various ducts, when vibrations are fed into the oval window, travelling waves along the basilar membrane are set up which cause the maximum displacement of the membrane at a place along the membrane which varies inversely as the logarithm of frequency. The theory is supported by experiments with models and observations on live guinea pig cochleas. Bekesy points out that, despite the differences between telephone, resonance, and travelling wave theories that are always emphasized, they are not so different when it comes to analysing transient sounds). The actual physical properties of the ear correspond to the travelling wave model, and travelling waves may be observed as transients in models set up to represent the basilar membrane, but there is the problem of being able to hear the "missing fundamental", which is considered to require some central analysis of nerve pulse repetition rate, since the pitch that is heard cannot be masked or caused to beat with other sounds. This would require activity of the telephone theory of hearing kind. See also *volley theory*.

trapezoid body

See *auditory pathways*.

transducer

A device to transform physical measurements such as pressure or temperature into electrical signals for purposes of processing information. Although it is a term intended for artificial devices, the human sensory organs are (by this definition) very high quality transducers that exceed the performance of any artificial devices ever built for similar conversion purposes.

transverse section

A section orthogonal to the axis from ventral to dorsal faces of an organism. A transverse section would divide a human into front and back portions.

trigram

Three letters appearing together, in order. A study of trigram frequencies may be of use in automated language processing. See *digram* and *tetragram*.

triphthong

A triphthong comprises three vowel articulations in succession. Like a diphthong, the vowels tend to be shortened and reduced, with an extended transition from one vowel articulation to the other. Common in British English (for example "fire" "f-ah-i-uh"). See also *diphthong*

tube model

A model of speech production that simulates the acoustic behaviour of the vocal tract using a transmission line (wave guide) model. Because the model can provide an accurate emulation of the real acoustic behaviour of the relevant acoustic tubes (pharynx, oral passage, and nasal passage), both the dynamics of change from posture to posture and the distribution of energy (especially in nasalised sounds) are more accurately modelled than in a conventional serial or parallel formant filter model. The control problem may be solved by using Carré's "Distinctive Region Model" (DRM) which is based on formant sensitivity analysis work carried out at the Speech Technology Laboratory of the KTH in Stockholm by Fant and his colleagues. The approach simplifies the control of the vocal tract to the manipulation of eight contiguous regions of unequal size. The approach is exploited by the author and his colleagues and is reported in several papers and reports (e.g. Real-time articulatory speech-synthesis-by-rules, and Low-level articulatory synthesis: a working text-to-speech solution and a linguistic tool)

unvoiced stop

See *voiced stop*.

uvula

See *velum*.

velocity of sound

See *speed of sound*.

velum

The extreme rear portion of the soft palate that terminates with the uvula (the bit that hangs down and can be seen at the back of the throat by looking into a mirror with the mouth open). The velum may be raised and lowered, opening and closing the connection between the main vocal passage and the nasal passage. The main vocal passage ultimately terminates at the mouth and the nasal passage at the nostrils.

vocal cords

See *vocal folds*.

vocal folds

More accurate name for vocal cords. The fleshy lips at either side of the glottis, within the larynx. They vibrate when tensioned and subject to enough airflow (from the lungs), giving rise to voiced speech sounds, and providing a basis for intonation patterns as the pitch may be changed by varying the tension. See glottis and voiced stop.

vocoder

A device for compressing the bandwidth of speech transmission by exploiting the redundancy inherent in the speech signal. There are as many different kinds of vocoder as there are ways of analysing speech. The first kind of vocoder built, and used for military security in the second world war, was the channel vocoder which exploits the fact that, in a spectrographic analysis, adjacent areas in the time-frequency-energy plot are highly correlated (or looking at it another way the output of a filter in a given frequency region only varies comparatively slowly -- say not faster than 100 hz, and the output is often quite similar to neighbouring filters). The speech is analysed into about 15 or 20 frequency channels and buzz/hiss detection and voicing detection are carried out to determine and transmit the excitation function. Just the energy per channel, the buzz/hiss state, and the voicing frequency are transmitted. The speech is then reconstituted using the channel signals to give the spectral envelope, and the buzz/hiss signal to determine the excitation type, with appropriate pitch (which also varies comparatively slowly) specified by the pitch frequency signal. A resonance vocoder carries out a similar excitation analysis but sends signals from a formant tracker to determine the spectral envelope. A resonance analog synthesiser then remakes the speech. There are many other kinds including several based on Linear Predictive Coding (LPC) analysis. The current state of the art achieves about 10 or 20:1 compression ratio. Many types suffer problems of tracking (pitch, formant frequency, etc.). LPC analysis currently provides the best basis for such speech compression schemes. LPC compression was used in the *Speak ‘n Spell* toy from Texas Instruments which came out in the summer of 1978. Because excessive compression was apparently used to meet cost targets for the intended market (roughly a 100:1 compression ratio, with the speech taking only 1200 bits per second) the speech quality left a something be desired, but was usable. The speech analysis was carried out “off-line” and the resulting compressed information stored in memory chips inside the machine, but the process was essentially the same as an LPC vocoder. It was an amazing advance in cost reduction for what had been until then an expensive military and research device.

voiced stop

A type of consonantal speech sound produced by completely closing off the vocal tract at some point, while keeping the velum also closed. Voiced stops are distinguished from unvoiced stops by the posture of the vocal folds. In voiced stops, the folds are adducted—held close together in the voicing position. In unvoiced stops, the vocal folds are abducted, opened wide. For this reason voicing starts much later for unvoiced than for voiced stops in moving into any following sound—an important acoustic cue. The transitions of formants are partly imposed on aspiration noise in unvoiced stops for some ten to thirty milliseconds, but on almost entirely on voiced spectral fine structure for voiced stops. If voicing continues during the closure of a voiced stop (as usually happens), then there is a pressure build up above the glottis, reducing the rate of air-flow through the vocal folds. This causes a drop of up to an octave in the instantaneous pitch and an associated rise in pitch upon release. Such small pitch movements are referred to as “micro-intonation” and are mainly associated with constriction of the vocal tract.

Other languages have finer distinctions, and divide stops into three categories based on the time of onset of voicing relative to the transitions associated with the release of the stop posture. See also *intonation*.

voicing

A kind of excitation of the vocal tract produced by forcing air through suitably adjusted vocal folds, causing them to vibrate fairly regularly in a frequency range stretching from 20 hz (at the end of utterances, for males) to 1000 hz (in a high note, for a female singer). The train of air pulses entering the vocal system has a harmonic energy distribution falling with frequency at about 6 db per octave. This spectrum is responsible for the spectral fine structure of voiced sounds which "carries" the resonant response of the vocal tract cavities (technically speaking, the spectrum of the voicing is convoluted with the spectral response of the vocal tract; both phenomena can be identified in the resulting speech spectrum). The spectral fine structure (voicing harmonics which occur at multiples of the voicing frequency) may be observed in a narrow band spectral analysis (say filter bandwidths of around 45 herz for male voices), which smears time variation and thus emphasizes the more slowly varying harmonic structure. The normal range of male spoken speech, neglecting glottal flap (creaky voice), is approximately 80 hz to 250 hz. The normal female range is roughly 30% higher on both limits due to a smaller larynx structure (the enlarged male larynx is a secondary sex characteristic). The range can be much greater in dramatic speech, and female singers can hit 1000 hertz (which leads to interesting problems since 1000 hertz is certainly higher than F1). See also *harmonics* and *source-filter model of speech production*.

voicing detection

The process of detecting by machine whether or not the vocal folds are vibrating during speech, on a moment to moment basis. The problem is difficult and not completely solved. It is not quite the same problem as pitch detection, though they are related. A good pitch detector would require a voicing detector component to avoid spurious pitch values being generated when voicing was absent. It is important to be able to detect the start and end of voicing accurately in all manner of speech processing systems, from bandwidth compressors to speech recognisers, since the distinction between a voiced stop and its voiceless counterpart, for example, may depend on ten to thirty milliseconds difference in the time of onset of voicing, compared to the timing of the formant transitions. With a "live" talker both the presence of voicing and its pitch may be determined fairly easily on a pitch-pulse to pitch-pulse basis using a laryngograph, or related equipment that measures the physical phenomena directly using high-frequency electrical resistance across the glottis, light transmission, or other directly related variable.

volley theory

The volley theory of hearing arises out of studies originated by Wever & Bray (1930) relating to the nerve impulses produced from the cochlea. The theory explains how nerve impulses occurring at frequencies exceeding 900 hz may be transmitted despite the 900 hz cap on individual fibres firing rates arising from the absolute refractory period which prevents a fibre from firing sooner than about a millisecond after a previous firing. At a particular phase of successive cycles of the stimulus sound pressure waveform, only some fraction of the potentially active nerve fibres fire, depending on their exact state of readiness, leaving others to fire during subsequent cycles. Experimental evidence supports the theory. With increasing frequency of stimulus up to 900 hz, the total intensity of nerve firing (pulses/msec) rises, and at 900 hz a further increase in frequency causes a drop in the density of nerve pulses to half what it was at 900 hz, but the inherent nerve pulse repetition rate continues to rise in phase with the stimulus frequency. At 1800 hz a similar effect occurs, with only one third of the nerves firing on any given cycle. This is consistent with the volley theory.

volume

The term volume denotes a quality of a sound that is distinct from loudness though both are subjective measures. Volume relates to the "size" or "extensity" of a sound. Loudness correlates best with acoustic intensity. Volume correlates better with the spread of associated neural activity. In *Equal volume judgements of tones,* Am. J. Psychology 62, 182-201, 1949, Thomas showed that if the total noise power of a band of continuous-spectrum noise is kept constant while the bandwidth is increased (thus activating a wider band on the basilar membrane), the volume increases more rapidly than the loudness. The figure shows the equal volume contours Thomas obtained. In addition to volume, loudness and pitch, sounds may differ from one another in terms of "brightness" (a triangle note sounds brighter than a wooden xylophone note), and in terms of "density" (compactness)and a trumpet note is harder, more compact, than the same note on the standard pipes of a pipe organ.

vowel

A speech sound produced using a relatively unobstructed vocal tract, and (except in whispered speech) with larynx excitation (voicing). Vowels are characterised by "quasi-steady-state" formant values, though in practice speech does not contain "steady states", from an acoustic point of view. From an articulatory point of view vowels are characterised by the position of the tongue hump (high/low, back/front, and degrees in between) that is associated with the notional vowel posture. In British English high front, low front, high back and low back extremes are exemplified by the vowel sounds in "heed", "had", "who'd", and "hard". John Holmes, of the Joint Speech Research Unit in England, showed that the notional vowel postures are by no means fixed, and vary greatly according to context. He displayed continuously varying F1/F2 values on a CRT. There were no obvious regions corresponding to the vowel articulations. Donald Broadbent, of the Cambridge University Psychology Department in England showed that the same word may be perceived as having a different vowel, depending on the carrier sentence in his famous "bit/bet/bat" experiment. The differences between the vowel phonemes of Educated Southern British English and those of General American are one of the important characteristics that distinguish the two accents The International Phonetics Association (IPA) has ratified a set of symbols to represent all kinds of speech sounds, but it is hard to quantify the unquantifiable. Even with the use of so-called cardinal vowels, which represent the extremes of vowel articulation (extreme high front, high back, low front, and low back vowels), together with numbered intermediate articulations, specifying the moving targets that comprise vowels is fraught with difficulty. See *vowel triangle.*

vowel triangle

If a plot of vowel sound is made against variation in the frequencies of formant 1 & 2 (F1 and F2), they are seen to form a roughly triangular figure. It is of interest that, correctly oriented, this figure also corresponds roughly to tongue hump positions required to produce the corresponding vowels. Some vowels (central vowels) fall inside the figure, as might be expected. The vowels and diphthongs represented in the diagrams are for General American. For the left-hand figure, clock-wise from the top left, the vowels are the ones in: "beat, bit, bet, bat, bart, bought, boot" with "book" and "but" for the left and right "inside" vowels respectively. For the right-hand figure, from top to bottom, the diphthongs are those in: "bate, bite, cute, quoit, bout, boat" respectively.

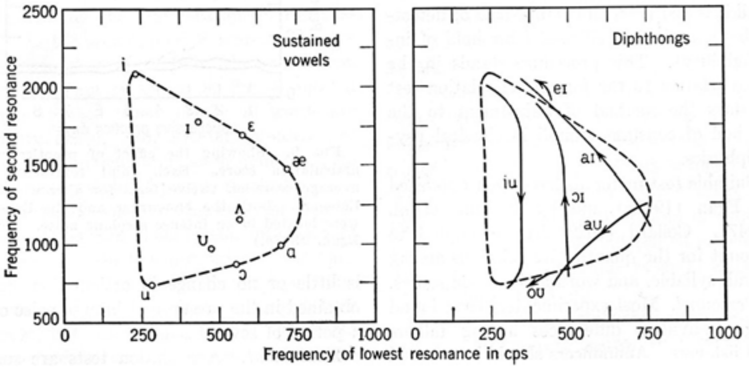

Figure 17: *The vowel triangle showing locations of vowels and diphthongs (Note: vowels & diphthongs are shown as IPA symbols)*

Page 1045 (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

VU meter

In transmitting or processing of the waveforms of speech and music there is a conflict between the most efficient use of the processing system and the fidelity of reproduction. If the system is set up to avoid overload on waveform peaks, then much of the time the depth of modulation, say on radio transmissions, may be less than half that possible. The signal to noise ratio, power requirements for given performance of transmitter, and range (which are inter-related) are likely to suffer. Automatic gain control is widely used to alleviate such problems but AGC circuits can overload so it is essential to be able to monitor the input and various stages of processing for maximum performance without degradation. It is also necessary to monitor subjective loudness (for the listener's benefit), which correlates with the waveform power rather than waveform peaks. Two classes of meter have developed for speech monitoring, those that respond to speech volume or energy and those that relate to speech peaks. The former are most widely implemented as the VU (or Volume Unit) meters used in America, and the latter as Peak Program Meters generally used throughout Europe and especially in the BBC. Both are calibrated in dbm (db re 1 milliwatt into 600 ohms). The real difference between properly designed meters of either type are (VUM versus PPM): (1) passive versus active; (2) load the line versus negligible load on the line; (3) different time constant definitions; (4) different scale markings; (5) hard to read versus easy to read.

watt

A measure of power generation or consumption rate. Energy flows whenever a force moves through a distance, including electrical forces in electrically resistive media and air pressure waves in air. Amounts of energy are equivalent regardless of how they are generated or consumed. To determine a total amount of energy generated or consumed, a time multiplier must be added. A watt is the rate at which energy is consumed when 1 amp of electrical current flows through 1 ohm or electrical resistance. The voltage (or electrical pressure) required to cause this to happen is 1 volt. A realistic measure of domestic electrical power consumption is the Kilowatt, and energy is metered as Kilowatt-hours. For small amounts of energy, energy consumption rates may be measured in milliwatts. One can talk about the energy flow per unit area. The threshold of human hearing corresponds to the ability to detect as little as 10 watts/m, which is equivalent to a milliwatt per cm2, or less than a milliwatt over the surface of the ear drum (as a continuous pressure fluctuation—a tone). It has also been stated in different terms—that the ear can detect movement nearly as small as the diameter of a hydrogen atom. The movement and the energy input must, in some sense, be equivalent.

waveguide

See *transmission line analog speech synthesiser*.

Weber's law

Also called the Weber/Fechner law. It states that, for a given stimulus dimension, the amount of increment (ΔI) needed to produce a just noticeable difference (JND) in the stimulus (I) is proportional to the stimulus intensity in that dimension. That is, ΔI/I = constant.

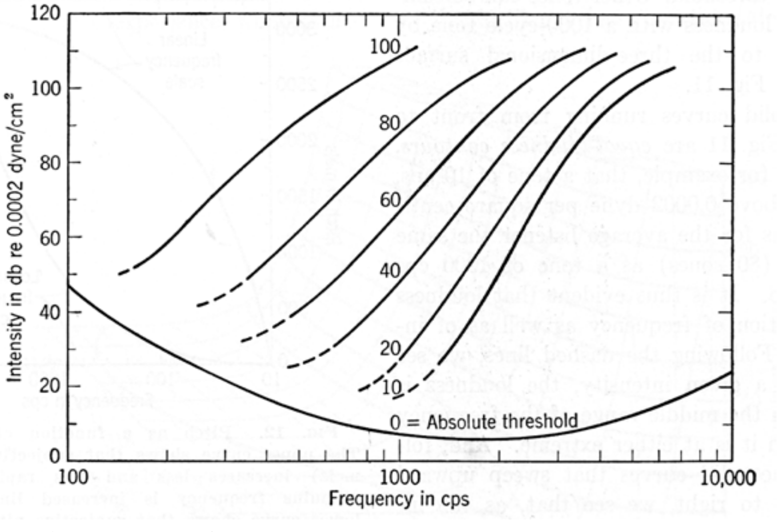

Figure 16: *Equal volume contours for pure tones*

Page 1004 (From Handbook of Experimental Psychology, edited by S.S. Stevens. © 1951 John Wiley & Sons. Used with permission of John Wiley & Sons)

WIP

Edinburgh University, Department of Applied and Theoretical Linguistics (as it now is) "Work in Progress"—was annual. There is now a web site for the department, and a web site for the Centre for Speech Technology Research (CSTR).

WPP

Phonetics Laboratory of the University of California of Los Angeles (UCLA) "Working Papers in Phonetics". Peter Ladefoged's laboratory. Their web site provides access to a photograph of Peter Ladefoged (no longer the director) teaching Rex Harrison and Wilfred Hyde-Whyte the niceties of Sweet's phonetic symbols on the set of My Fair Lady, the classic adaptation of George Bernard Shaw's Pygmalion. The lab is now part of the interdisciplinary Speech Processing & Auditory Perception Laboratory at UCLA.

zero-crossing analysis

A speech waveform is bi-polar, crossing and recrossing the zero pressure reference corresponding to the prevailing static pressure (atmospheric pressure). The rate at which these crossings occur is called the zero-crossing rate and will be proportional to the frequency content of the speech in a rather complicated fashion. The greater and the higher the high frequency content is, the higher will be the zero-crossing rate. A more accurate idea of the waveform, though still not bearing a simple relationship to the frequency components, is the so-called zero-crossing interval analysis, which generates a histogram of the frequency of occurrence of intervals of different duration between successive zero-crossings. The value of this representation is improved by differentiating the speech before processing it. Since the differentiation effectively makes what were local peaks in the original waveform (or "turn-arounds") into zero-crossings, a zero-crossing interval analysis of speech which has been differentiated (and a telephone carbon transmitter does this for nothing!) is exactly the same as an analysis of turnaround intervals. In either case the higher frequency components are better represented, as might be expected. A zero-crossing interval histogram is reminiscent of a frequency spectrum, though different. A zero-crossing analysis of an unprocessed speech waveform misses many of the higher frequency components.